

AD-A060 115

SYSTEM DEVELOPMENT CORP SANTA MONICA CALIF
FACP SPEECH RECOGNITION/TRANSMISSION SYSTEM.(U)
AUG 78 B T OSHIKA

F/G 17/2

UNCLASSIFIED

SDC-TM-6144/000/00

RADC-TR-78-193

F30602-77-C-0056

NL

1 OF 1
AD
A060115



AD A060115

DDC FILE COPY

② LEVEL II

RADC-TR-78-193
Final Technical Report
August 1978



FACP SPEECH RECOGNITION/TRANSMISSION SYSTEM

B. T. Oshika

System Development Corporation

Approved for public release; distribution unlimited.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

DDC
RECEIVED
OCT 18 1978
B

78 10 10 106

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-78-193 has been reviewed and is approved for publication.

APPROVED:

Melvin G. Manor
MELVIN G. MANOR
Project Engineer

APPROVED:

Howard Davis
HOWARD DAVIS
Technical Director
Intelligence and Reconnaissance Division

FOR THE COMMANDER:

John P. Huss
JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA), Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return this copy. Retain or destroy.

MISSION
of
Rome Air Development Center

RADC plans and conducts research, exploratory and advanced development programs in command, control, and communications (C³) activities, and in the C³ areas of information sciences and intelligence. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19 REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-78-193	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) FACP SPEECH RECOGNITION/TRANSMISSION SYSTEM	5. TYPE OF REPORT & PERIOD COVERED Final Technical Report April 10 77 - May 19 78	6. PERFORMING ORG. REPORT NUMBER TM-6144/000/00
7. AUTHOR(s) B. T. Oshika	8. CONTRACT OR GRANT NUMBER(s) F30602-77-C-0056	
9. PERFORMING ORGANIZATION NAME AND ADDRESS System Development Corporation 2500 Colorado Avenue Santa Monica CA 90406	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62702F 55561231	
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAA) Griffiss AFB NY 13441	12. REPORT DATE August 1978	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Same	13. NUMBER OF PAGES 63	
	15. SECURITY CLASS. (of this report) UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) Same		
18. SUPPLEMENTARY NOTES RADC Project Engineer Melvin G. Mauor, Jr. (IRAA)		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) acoustic-phonetics linear predictive coding speech compression area functions narrowband transmission speech recognition dyad synthesis phoneme vocoder speech transmission fundamental frequency speech analysis vocoder		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report describes a phoneme vocoder capable of transmitting compressed speech data over bandlimited communication channels at rates lower than 200 bits per second. Using linear prediction analysis for parameter extraction, and sophisticated segmentation and labeling techniques, the vocoder analyzer codes the incoming speech signal into a sequence of discrete sound units, or phonemes. At the receiving end of the channel, the phoneme sequence is input to a digital speech synthesizer. An area function dyad synthesis procedure (Cont'd)		

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DDC
RECEIVED
OCT 18 1978
B

339900
78 10 10 100
JP

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20 (Cont'd)

is described which is based on an area function model representing the vocal tract as a set of 14-cross-sectional areas. Area functions representing phoneme steady states (nuclei) and transitions between any ordered pair of phonemes (dyads) are stored in a dyad table. Given an input phoneme string, the synthesizer selects the corresponding sequence of nuclei and transitions and interpolates between each of the 14 cross-sectional areas, producing a model of the shape of the vocal tract changing in time. The filtering process of this vocal tract model is identical to the optimum inverse filter of linear prediction analysis, allowing direct conversion to linear predictive coding (LPC) synthesis. A terminal analog synthesizer is also described. ~~A terminal analog synthesizer is also described.~~ Diagnostic Rhyme Tests (DRT) of vocoder performance for two male speakers yielded scores of 70.6% using area function dyad synthesis and 83.5% for terminal analog synthesis.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)


CONTENTS

1. INTRODUCTION	1
2. BACKGROUND	2
3. ANALYSIS	5
3.1 The A-Matrix	5
3.2 Fundamental Frequency Estimation	5
3.3 Spectral Analysis	8
3.4 Phoneme Segmentation and Labelling	11
3.5 Prosodic Analysis and Label Smoothing	17
4. SYNTHESIS	19
4.1 Area Function Dyad Synthesis	19
4.2 Terminal Analog Synthesis	25
5. TESTING	28
6. EVALUATION	31
6.1 Condition 1	31
6.2 Condition 2	37
6.3 Conditions 3 and 4	42
6.4 Diagnostic Patterns	51
7. SUMMARY AND RECOMMENDATIONS	53
8. REFERENCES	55

ACCESSION for	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Buff Section <input type="checkbox"/>
UNANNOUNCED <input type="checkbox"/>	
JUSTIFICATION _____	
BY _____	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL. and/or SPECIAL
A	

EVALUATION

The objective of this effort was to design and develop a narrowband voice transmission capability to operate in conjunction with Tactical Command and Control Communications for use on the FACP Speech Recognition/Transmission System. The feasibility of transmitting compressed speech data at rates lower than 200 bits per second has been demonstrated. This work is in support of TPO 3B.


MELVIN G. MANOR, JR.
Project Engineer

1. INTRODUCTION

The objective of the research effort was to design and develop a narrowband voice transmission capability to operate in conjunction with Tactical Command and Control Communications, for use on the FACP Speech Recognition/Transmission system.

To achieve this objective, SDC developed a phoneme vocoder (voice coder), capable of transmitting compressed speech data over band-limited communication channels at rates lower than 200 bps. This low bit rate is possible because the incoming speech signal is coded by an analyzer into a sequence of a relatively small number of discrete sound units, or phonemes, which carry maximum phonological information.

At the receiving end of the communication channel, the phoneme sequence is input to a digital speech synthesizer which transforms the sequence into acoustic speech parameters. Much of the SDC research effort has focused on defining the complex transformations needed to convert phoneme strings into intelligible speech. In particular, SDC has developed an area function dyad synthesizer based on area functions which model the changing shape of the vocal tract during speech production.

This report includes background on speech/analysis systems, descriptions of the SDC analyzer and area function dyad synthesizer, discussion of phoneme discriminability tests and results, and recommendations for future research.

2. BACKGROUND

Various approaches to analysis/synthesis systems have been reviewed elsewhere [Flanagan 1965, Schroeder 1966, Gold 1977] and will be only briefly discussed here.

The simplified model of speech production used in almost all analysis/synthesis systems assumes that the vocal tract is a non-uniform acoustic tube not coupled with the nasal cavity. The excitation function is separated from the vocal tract transfer function which is considered as a slowly time-varying linear filter. Many vocoder systems also assume that excitation is periodic or noiselike, but not both.

In general, at the transmission end of a vocoder system, the analyzer must determine whether the excitation is periodic or noiselike (voicing detection), estimate the fundamental frequency of the periodic excitation (pitch extraction), and estimate the spectral envelope representing the vocal tract shape. Because of the non-stationary properties of speech, analysis is usually done on a frame-by-frame basis.

At the receiving end, the synthesizer must reconstruct the spectral information and apply the appropriate excitation function to produce intelligible speech which has a spectrum approximating the spectrum of the original speech.

The channel vocoder, originally developed in the 1930's, uses a bank of band pass filters, rectifiers and low-pass filters to estimate the short-time amplitude spectrum [Dudley 1939, Gold and Rader 1967]. At the receiver, appropriate excitation is applied to a corresponding bank of band pass filters and modulators to reconstruct a short-time spectrum approximating that of the original speech. Typical present day implementations of channel vocoders require about 2400 bps for transmission of reasonable quality speech.

Another currently popular class of vocoders is based on linear predictive coding (LPC) which assumes that a sample of a speech waveform can be predicted as a linear weighted sum of previous samples [Saito and Itakura 1966, Atal and Schroeder 1967, Atal and Hanauer 1971, Makhoul 1975, Markel and Gray 1976]. In LPC vocoders, linear prediction coefficients are extracted for each frame of speech data to obtain an inverse model of the speech spectrum. Voicing parameters, pitch and gain are also computed. At the receiver, an excitation signal is constructed from the voicing and pitch information and drives a synthesis filter corresponding to the inverse of the analysis model. The gain function is used to match the energy of the synthetic speech to that of the original speech. Very high quality speech can be obtained down to about 3,300 bps, with some degradation down to about 1,400 bps, and complete degradation below 1,400 bps [Markel and Gray 1976].

Other types of speech encoding-decoding systems include formant vocoders [Flanagan et al. 1962], maximum likelihood vocoders [Itakura and Saito 1968], and homomorphic vocoders [Oppenheim 1969]. In general, these systems require data rates of several thousand bps for acceptable quality.

SDC has attempted to significantly reduce the required bandwidth for intelligible speech transmission by coding the acoustic parameter output of the analyzer in terms of phonemes, i.e., the alphabet of distinctive speech sounds of a language. Spoken English can be described by a set of about 48 phonemes which can be encoded in 6 bits. If quantized pitch (3 bits) and duration (2 bits) for each phoneme are included, then each phoneme can be represented by 11 bits. For an average speaking rate of 12 phonemes per second, a transmission rate of 132 bps is possible. In such a phoneme vocoder, the analyzer must not only extract relevant acoustic parameters from the speech signal, but must also classify the parameters into a sequence of sound

segments with accurate phoneme labels. The synthesizer must accept a phoneme string as input, and generate corresponding acoustic parameters.

In the SDC phoneme vocoder, the analyzer uses linear prediction for spectral analysis and a fast frequency domain pitch algorithm for fundamental frequency estimation [Gillmann 1975]. The analysis procedures and segmentation and labeling schemes are described in detail in Section 3.

The synthesizer is based on an area function model which represents the vocal tract as a set of 14 cross-sectional areas. A 48 x 48 table describes 14 cross-sectional areas for each of the 48 phonemes. The steady states (nuclei) of each phoneme are represented down the main diagonal of the table, and the transition states (dyads) between any ordered pair of phonemes are the off-diagonal entries. Given a phoneme string as input, the synthesizer selects the corresponding sequence of nucleus and dyad entries from the table and interpolates between each of 14 cross-sections, producing a model of the shape of the vocal tract changing in time.

It has been shown that the filtering process of this vocal tract model is identical to the optimum inverse filter of linear prediction analysis [Wakita 1972], thus allowing direct conversion to LPC synthesis from the vocal tract model. The synthesizer is described in greater detail in Section 4.

3. ANALYSIS

3.1 The A-Matrix

The analyzer, running on a Raytheon 704 minicomputer, is based on an acoustic-phonetic processor developed by SDC during an extensive research and development program in speech understanding research [Bernstein 1975, Ritea 1975]. The analyzer accepts, digitizes and records speech input in real time. Subsequent processing in non-real time extracts acoustic parameters for each 10 msec. segment of speech, such as fundamental frequency and formant frequencies and amplitudes, and uses these parameters to assign a sequence of phoneme labels to the speech. The parametric information and phoneme choices are stored in the A-matrix, summarized in Table 3-1. The components of the analyzer are shown in Figure 3-1.

Input speech is low-pass filtered and digitized at a rate of 20,000 samples per second by a 12-bit analog-to-digital converter.

A smoothed root mean square (RMS) energy value is then calculated for each 10 msec. frame of speech and silence areas are marked.

3.2 Fundamental Frequency Estimation

Fundamental frequency (F0), (or pitch), is estimated using a fast frequency domain pitch algorithm developed by Gillmann [1975], which operates in three phases:

1. Down-sampling. A digital filter is used to down-sample the digitized speech from 20,000 to 2,000 samples per second, to remove frequencies that lie outside the range of possible fundamentals and thus reduce computation time.

Table 3-1. A-Matrix Contents

- Rough labels: silence, sonorant, unvoiced, etc.
- Smoothed RMS
- F0: pitch track
- Voicing indicators: fry, falsetto, sporadic, etc.
- F1-4: formant frequencies and amplitudes
- Formant discontinuity indicators
- F2: direction-of-change indicator
- LPC amplitudes: first point, first two peaks
- Energy functions: sonorant, low-frequency, edge-frequency
- Sonorant indicators: retroflex, lateral, nasal
- Boundaries: phone, syllable, phrase
- Phrase type: falling, rise-fall, etc.
- Vowel stress
- Rate of speech
- Slope change count
- FRIC/PLOS indicators: closure, (voiced) burst, spectral tags, etc.
- Phone labels and scores

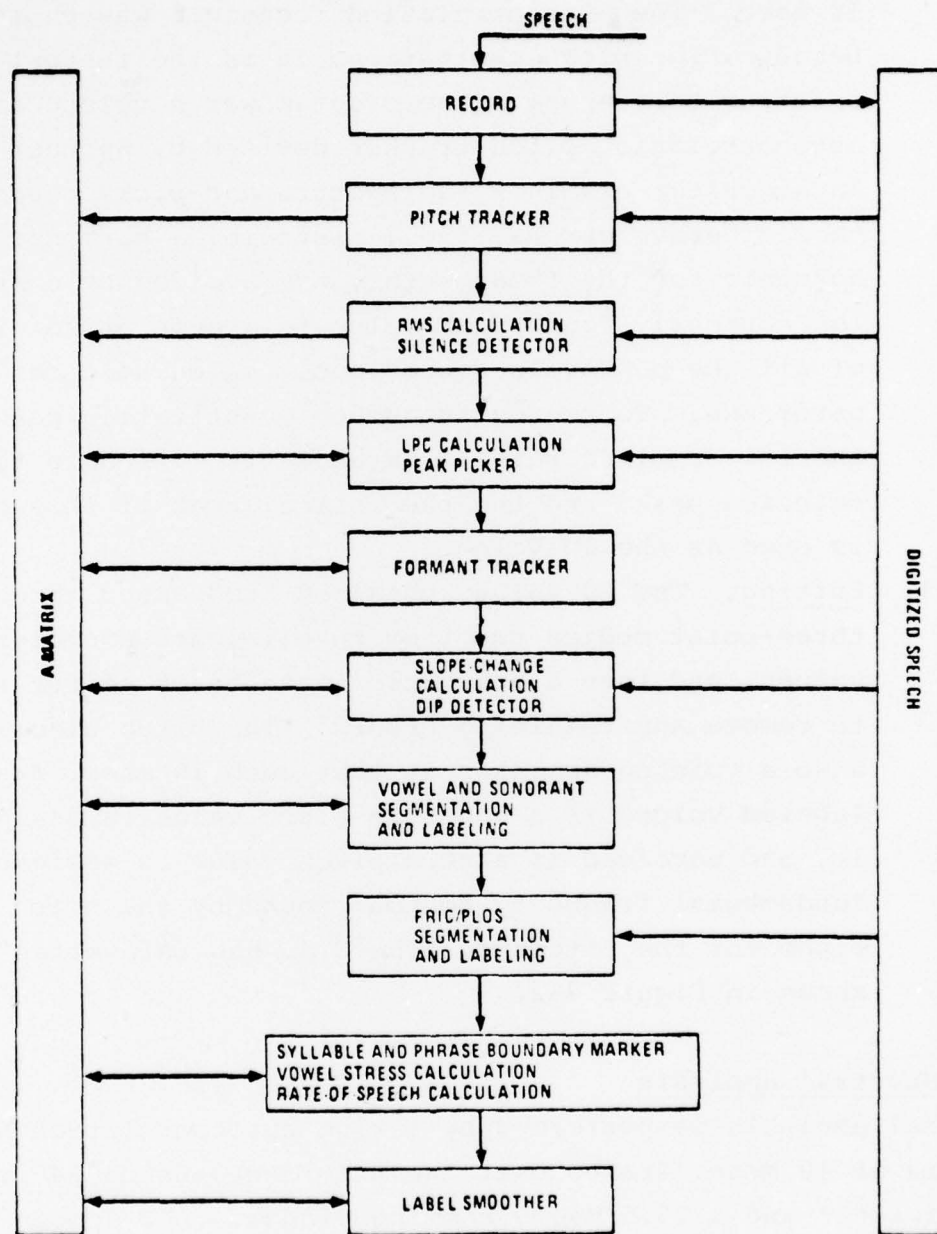


Figure 3-1. Analyzer

2. Autocorrelation and pitch extraction. An autocorrelation spectrum with a window size of 50 msec. is taken every 10 msec. The autocorrelation technique was chosen because, for pitch estimation, it is the fastest spectrum to compute. The program was developed from an autocorrelation pitch tracker devised by Skinner [1975]. An algorithm examines the spectra and picks peaks from them. Octave errors, i.e., mistaking a harmonic or sub-harmonic for the fundamental, are avoided by comparing the currently found peak value to a cumulative average of all the peak values previously calculated for the utterance. To reduce frequency quantization caused by the low sampling rate, a parabola is fitted to the selected peak, and the theoretical peak of this parabola is used as the FO value.
3. Editing. The FO values obtained are passed through a three-point median smoother to eliminate anomalous values, and then a heuristic pitch track editor attempts to remove any remaining errors. The pitch algorithm is also a voicing detector in that each 10 msec. frame is labeled voiced if a non-zero pitch value is assigned to it, and unvoiced if a zero pitch value is assigned. The fundamental frequency contour found by the pitch algorithm for the utterance "The U.S. has Lafayettes" is shown in Figure 3-2.

3.3 Spectral Analysis

Spectral analysis is performed by taking autocorrelation LPC spectra at 10 msec. frames over the utterance, using 24 predictor coefficients and a 25.6 msec. Hamming window. LPC speech processing techniques assume that vocal tract resonance characteristics of voiced speech can be modeled by an all-pole filter of the form

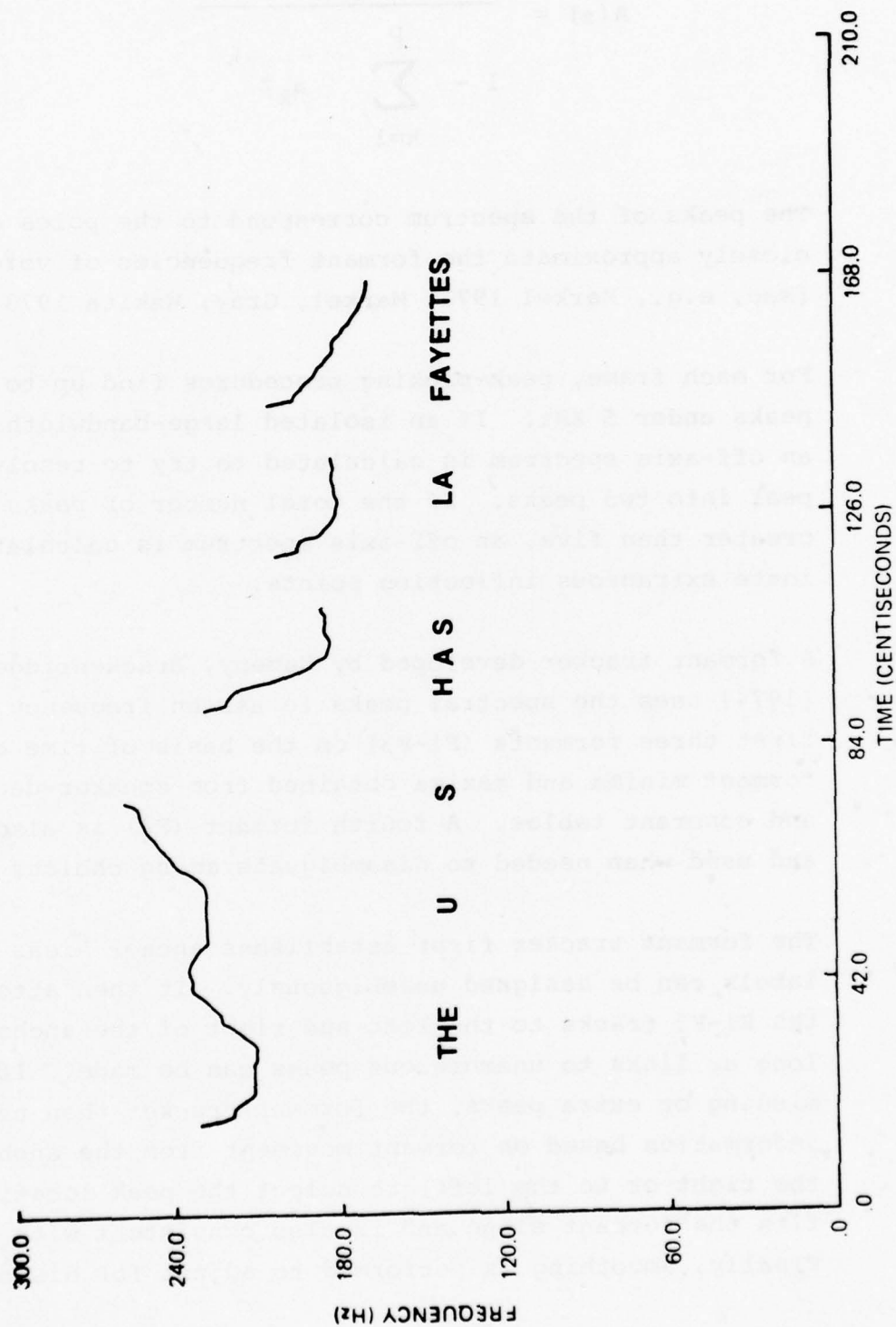


Figure 3-2. Fundamental Frequency Contour of the Utterance "The U.S. has Lafayette's."

$$A(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}}$$

The peaks of the spectrum correspond to the poles of $A(z)$ and closely approximate the formant frequencies of voiced sounds [see, e.g., Markel 1971, Markel, Gray, Wakita 1973].

For each frame, peak-picking procedures find up to five frequency peaks under 5 KHz. If an isolated large-bandwidth peak is found, an off-axis spectrum is calculated to try to resolve the broad peak into two peaks. If the total number of peaks found is greater than five, an off-axis spectrum is calculated to eliminate extraneous inflection points.

A formant tracker developed by Kameny, Brackenridge and Gillmann [1974] uses the spectral peaks to assign frequency values to the first three formants (F1-F3) on the basis of time constraints and formant minima and maxima obtained from speaker-dependent vowel and sonorant tables. A fourth formant (F4) is also calculated and used when needed to disambiguate among choices for F1-F3.

The formant tracker first establishes anchor areas in which F1-F3 labels can be assigned unambiguously. It then attempts to extend the F1-F3 tracks to the left and right of the anchor areas, as long as links to unambiguous peaks can be made. If there are missing or extra peaks, the formant tracker then uses slope information based on formant movement from the anchor (either to the right or to the left) to select the peak location that best fits the formant slope and is also consistent with adjacent peaks. Finally, smoothing is performed to adjust for discontinuities.

Following the formant-tracking pass, the number of slope changes (in the digitized speech) per 10 msec. frame is calculated and dip areas are marked using an algorithm similar to that described by Weinstein et al. [1975]. Dip information is used to enhance detection of plosives, nasals, and flap gestures as well as boundaries. Figure 3-3 graphically shows the results of processing an utterance to this point.

3.4 Phoneme Segmentation and Labeling

Two distinct algorithms are used for segmentation and labeling, one for classification of vowels and sonorants, and one for fricatives and plosives. The English vowels and sonorants, listed in Table 3-2, are voiced sounds made with relatively unobstructed passage of air through the vocal tract, and are characterized by fairly clear formant patterns. Fricatives, on the other hand, are made with air passing through a constriction, causing turbulence. Plosives are characterized by complete closure followed by a burst. The fricatives and plosives of English are shown in Table 3-3.

The segmentation and labeling techniques can use up to three levels of labels: (1) phoneme symbols as shown in Tables 3-2 and 3-3, (2) feature bits indicating phonetic modifications such as nasalization and retroflexion, and (3) rough labels indicating broad categories such as "voiced" or "vowel."

Vowel-sonorant analysis

For vowel-sonorant identification [Kameny 1976], F1-F3 values are converted to linear scale values that correspond to scaled values in speaker-dependent vowel-sonorant table. Phoneme labeling is done on the basis of linear distance, so that a movement of \underline{n} units in any formant is equivalent to a movement of \underline{n} units in another formant.

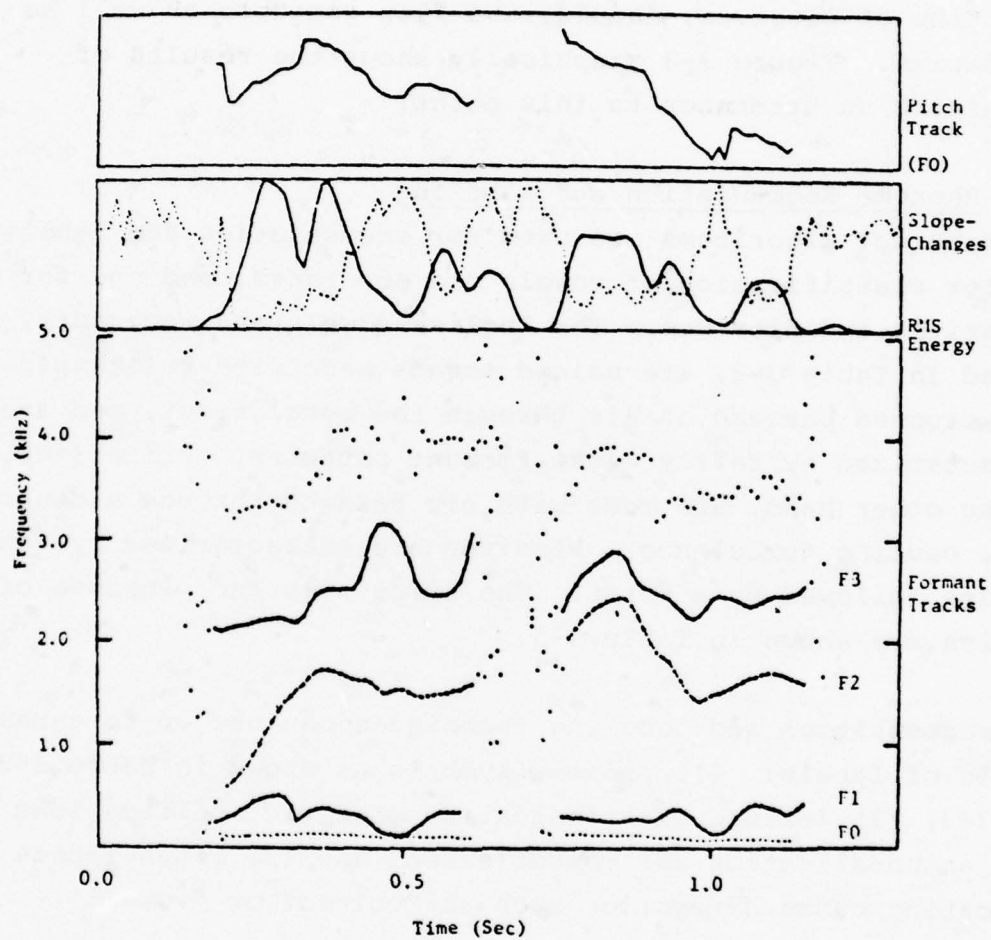


Figure 3-3. Graph of the Parameters for the Utterance
"What is the speed of it?"

Table 3-2. Vowels & Sonorants

<u>Phoneme</u>	<u>Machine Representation</u>	<u>Example</u>
i	IY	beat
I	IH	bit
e	EY	bait
ɛ	EH	bet
æ	AE	bat
ɑ	AA	Bob
ʌ	AH	but
ɔ	AO	bought
o	OW	boat
U	UH	look
u	UW	boot
ə	AX	about
ɹ	IX	roses
ɜ	ER	bird
aU	AW	down
aI	AY	buy
ɔI	OY	boy
y	Y	you
w	W	wit
r	R	rent
l	L	let
m	M	met
n	N	net
ŋ	NX	sing
ʍ	WH	which
l̥	EL	battle
m̥	EM	bottom
n̥	EN	button
r̥	DX	batter

Table 3-3. Fricatives & Plosives

<u>Phoneme</u>	<u>Machine Representation</u>	<u>Example</u>
p	P	<u>p</u> et
t	T	<u>t</u> en
k	K	<u>k</u> it
b	B	<u>b</u> et
d	D	<u>d</u> ebt
g	G	<u>g</u> et
h	HH	<u>h</u> at
f	F'	<u>f</u> at
θ	TH	<u>th</u> ing
s	S	<u>s</u> at
ʃ	SH	<u>sh</u> ut
v	V	<u>v</u> at
ð	DH	<u>th</u> at
z	Z	<u>z</u> oo
ʒ	ZH	a <u>z</u> ure
ç	CH	<u>ch</u> urch
ʝ	JH	<u>j</u> udge

Segmentation begins by locating phoneme nuclei in areas that are voiced, have F1-F3 values, do not contain local RMS energy dips, have slope-change counts below a threshold, and are at least 40 msec. long. Within such a voiced area, the nucleus finder first locates the frame(s) of peak RMS energy. It also calculates the absolute first-difference in scaled F1-F3 values in adjacent frames, and if this value exceeds a threshold, no nucleus is found because formant frequencies are changing too rapidly to define a steady-state. If the values lie below a threshold, the frame with the minimum difference is chosen as the nucleus. If there are several such frames, the one closest to the RMS peak is selected.

Segment boundaries are defined by moving in both directions from the nucleus until scaled F1, F2 or F3 values for more than a single frame differs from that of the nucleus by more than a threshold, or until a previously defined boundary of an adjacent segment is found.

Label choices and scores are determined by computing linear distances from scaled F1-F3 values of the unknown nucleus to values for each vowel and sonorant stored in the speaker table. The labels of the first four closest matches below a threshold are chosen, along with scores which are the linear distance between the F1-F3 values of the nucleus and the F1-F3 values of the selected stored targets. Contextual information, such as the possible influence of an adjacent nasal or retroflex sound, are taken into account in selecting labels.

Fricatives and plosives

The fricative and plosive algorithm [Molho 1976] uses the autocorrelation method of linear prediction for spectral analysis, but with only 8 coefficients and a narrower (6 msec.) Hamming window than used in vowel-sonorant identification. Parameters extracted from the spectra are:

FRL: lowest frequency peak
FRS: second-lowest frequency peak
DBM: maximum spectral amplitude
SHR: a sharpness measure of FRL
FRA: a measure of spectral skew
BFRL: a bit which is set if the amplitude of the FRL is
within 6 dB of DBM
BFRS: a corresponding bit for the FRS peak

Parameters are extracted for every unvoiced 10 msec. frame, and also for frames where there is transition into and out of voicing, a missing first formant, a zero-crossing count exceeding a threshold in a voiced area (typical of voiced strident fricatives such as /z/), or indication of an unexplained energy dip pattern determined by the vowel-sonorant analysis.

A fricative segmentation process groups together unvoiced frames on the basis of frame-to-frame stability constraints on parameters DBM and FRA, and tries to assign preliminary phoneme labels S, SH and/or HH on the basis of average parameter values within the group.

At this point segments are merely concatenated groups with the same label, and some segments or parts of segments may overlap and have up to three labels. Duration rules are applied to remove obviously mislabeled short segments, and remaining segments are assigned preliminary labels and scores.

Plosives are now segmented, with the frame having the minimum DBM used as a reference for identifying silence in the utterance. A burst-marking routine identifies up to four non-silent frames preceded by silence as potential plosive bursts. A closure algorithm compares F1 motion in the vicinity of unvoiced areas to a hypothetical linear formant transition between surrounding vowel nuclei, and oral closure is identified if F1 is

sufficiently low. Parameters for plosive bursts are then examined, including spectral parameters up to 40 msec. after burst onset, voice onset time, F2 and F3 values at voice onset, and contextual information such as existence of adjacent S. Plosives are then labeled and scored on the basis of patterns of burst parameters, and plosive aspiration originally labeled as HH by the fricative analysis is removed.

Final fricative labels and scores are now assigned on the basis of adjacent voicing, oral closure, duration, and the previously determined contextual information.

3.5 Prosodic Analysis and Label Smoothing

In addition to phoneme segmentation and labeling, prosodic characteristics such as pitch contours, linguistic stress, rate of speech, and syllable and phrase boundaries are determined. Pitch contours at phrase and utterance boundaries are labeled as falling, rising, rise-fall, etc. Intensity, duration and pitch are used to determine levels of vowel stress. Rate of speech is calculated using a 1 second window. A "convex hull" algorithm based on a loudness function is used to locate syllable and phrase boundaries [Mermelstein 1975].

On its final pass the analyzer makes a "best" label choice for each 10 msec. frame of speech and enters it in the A-matrix. These labels are then corrected by a symbolic smoothing algorithm. An example of a portion of a final A-matrix is shown in Figure 3-4.

10-MSEC. SEGMENT NUMBER																	
ROUGH SEGMENT LABEL																	
RMS ENERGY																	
FUNDAMENTAL FREQUENCY																	
VOWEL-SONORANT CHOICES 1,2,3&4 (WHERE APPLICABLE)																	
FRICATIVE-PLOSIVE CHOICES 1,2&3 (WHERE APPLICABLE)																	
FORMANT FREQUENCIES 1,2,3&4																	
STRESS																	
VOCODER LABEL																	
22	SO	54	109	W18	NA50	0	0	0	0	0	0	293	585	2113	3279	S	W
23	SO	87	109	W18	NA50	0	0	0	0	0	0	322	647	2138	3203	S	W
24	V	125	111	0	0	0	0	0	0	0	0	359	710	2146	3269	S	W
25	V	171	114	0	0	0	0	0	0	0	0	401	787	2158	3283	S	W
26	V	228	117	0	0	0	0	0	0	0	0	439	886	2175	3314	S	AX
27	VW	280	117	AX	2	OW15	UH19	AH29	0	0	0	466	997	2192	3309	S	AX
28	VW	302	119	AX	2	OW15	UH19	AH29	0	0	0	486	1112	2208	3344	S	AX
29	VW	296	121	AX	2	OW15	UH19	AH29	0	0	0	500	1224	2218	3360	S	AX
30	VW	288	123	AX	2	OW15	UH19	AH29	0	0	0	505	1328	2223	3332	S	AX
31	VW	274	124	AX	2	OW15	UH19	AH29	0	0	0	485	1412	2219	3320	S	AX
32	V	220	124	0	0	0	0	0	0	0	0	406	1470	2196	3405	S	AX
33	V	147	126	0	0	0	0	DX50	B55	DH55	262	1519	2181	3846	S	DX	
34	V	126	137	0	0	0	0	DX50	B55	DH55	309	1581	2237	3179	S	DX	
35	V	177	137	0	0	0	0	DX50	B55	DH55	352	1649	2354	3546	S	DX	
36	VW	247	137	IH16	IX18	UH34	0	0	0	0	0	379	1701	2448	3634	S	IH
37	VW	291	135	IH16	IX18	UH34	0	0	0	0	0	386	1726	2487	3750	S	IH
38	VW	298	135	IH16	IX18	UH34	0	0	0	0	0	380	1720	2500	3593	S	IH
39	VW	271	133	IH16	IX18	UH34	0	0	0	0	0	368	1712	2507	4064	S	IH
40	VW	232	131	IH16	IX18	UH34	0	0	0	0	0	353	1687	2510	4070	S	IH
41	VW	197	127	IH16	IX18	UH34	0	0	0	0	0	338	1659	2503	4082	S	IH
42	VW	161	124	IH16	IX18	UH34	0	0	0	0	0	326	1638	2511	3984	S	IH
43	V	124	123	0	0	0	0	0	0	0	0	313	1629	2599	3616	S	IH
44	V	99	121	0	0	0	0	Z25	L	0	286	1624	2798	3798	S	Z	
45	V	87	116	0	0	0	0	Z25	0	0	238	1602	3019	4103	S	Z	
46	V	78	113	0	0	0	0	Z25	0	0	186	1549	3142	3988	S	Z	
47	V	68	111	0	0	0	0	Z25	0	0	152	1506	3157	0	S	Z	
48	V	55	110	0	0	0	0	Z25	0	0	137	1512	3125	0	S	Z	
49	V	41	110	0	0	0	0	Z25	V50	DH55	130	1537	3073	0	S	Z	

Figure 3-4. Portion of the A-Matrix for the Utterance
"What is the speed of it?"

4. SYNTHESIS

During the development of the SDC phoneme vocoder, much of the research focused on design and implementation of an area function dyad synthesizer based on interpolation of the pre-stored cross-sectional areas of the vocal tract. Some modifications were also made to an existing digital simulation of a terminal analog synthesizer involving manipulation of acoustic parameters by rule.

4.1 Area Function Dyad Synthesis

The area function approach to synthesis is based on an acoustic tube model of the vocal tract directly derivable from the speech waveform given the following assumptions [Atal and Hanauer 1971, Wakita 1972]:

- 1) Although the vocal tract is a non-uniform tube, it can be adequately represented as a set of M connected cylindrical sections of equal length, each section having uniform cross-sectional area.
- 2) Sound propagation through each area can be treated as a plane wave, and assumptions associated with elementary wave propagation are valid.
- 3) Sections are rigid and losses due to factors such as wall vibration or viscosity can be ignored.
- 4) The model is linear and not coupled to the glottis, and does not consider effects of the nasal cavity.

Atal [1970] has shown that formant frequencies and bandwidths are sufficient to uniquely determine the areas of an acoustic tube made up of a specified number of sections, and demonstrated that a transfer function with M poles can always be realized as the transfer function of an acoustic tube of M cylindrical sections. Wakita [1972] showed that the filtering process of the acoustic tube model of the vocal tract is identical with the optimum inverse filter of linear prediction analysis. This means that the reflection coefficients, which uniquely define

the area ratios of the acoustic tube model, can be obtained directly by linear prediction analysis, and therefore vocal tract shapes can also be estimated directly.

From the point of view of synthesis, it means that given a specified vocal tract shape, a corresponding LPC synthesis filter can be easily constructed. This fits well into a phoneme vocoder scheme, because phoneme choices which are the output of the analyzer can be associated with distinctive vocal tract shapes in the articulatory domain.

An important advantage of using area functions for phoneme synthesis is that cross-sectional areas of the vocal tract can be assumed to vary in a relatively slow and continuous way in time because of the dynamics of the tongue and other articulators. Therefore, interpolation of pre-stored areas should yield a better representation of speech production than interpolation of other pre-stored filter parameters such as LPC parameters or reflection coefficients. Experiments showed that synthetic speech produced by interpolation of LPC parameters and of reflection coefficients was characterized by popping and clicking sounds, indicating abrupt discontinuities. Therefore, synthesis based on area functions was selected.

The area function synthesizer can be viewed as an $n \times n$ dyad table describing M representative cross-sectional areas of the vocal tract for each of n phonemes. The steady states, or nuclei, of the phonemes are the main diagonal entries, and transition states between any ordered pair of phonemes, or dyads, are the off-diagonal entries. For example, in the 2×2 table shown in Figure 4-1, IY and T are phoneme names, diagonal entries $IYIY$ and TT are nuclei, and IYT and TIY are transitions. Stored for each nucleus or transition are 14 cross-sectional areas, A_1 - A_{14} , specifying a representative vocal tract shape.

Phoneme Symbols	IY	T
IY	A1-A14	A1-A14
T	A1-A14	A1-A14

Figure 4-1. Example of a Two Phoneme Dyad Table

The current version of the synthesizer uses M=14 cross-sectional areas and n=48 phonemes. However, any one or two character symbol can be entered in the table as the name of a sound unit, so that more detailed phonetic units can be used in addition to the standard phoneme alphabet.

The set of area functions which represents a nucleus or transition is selected manually from analysis of citation words recorded by a single male speaker. A linear prediction analysis is performed on the digitized utterances and 14 cross-sectional areas calculated for each 10 msec. frame of speech. Figure 4-2 shows the area functions for the word "gob." Areas of sections closer to the glottis are graphed close together because they contribute relatively little to distinctive vocal tract shape. Areas of sections closer to the lips are graphed further apart so that changes in area can be more clearly seen.

Speech frames which represent steady states or transitions, such as those marked in Figure 4-2, are chosen using a combination of listening and inspection of the area functions and corresponding waveform, available through an interactive speech laboratory facility. The name of the nucleus or transition (e.g., IYIY or IYT) is entered into the dyad table along with the set of area functions A1-A14 corresponding to the selected frame. The name

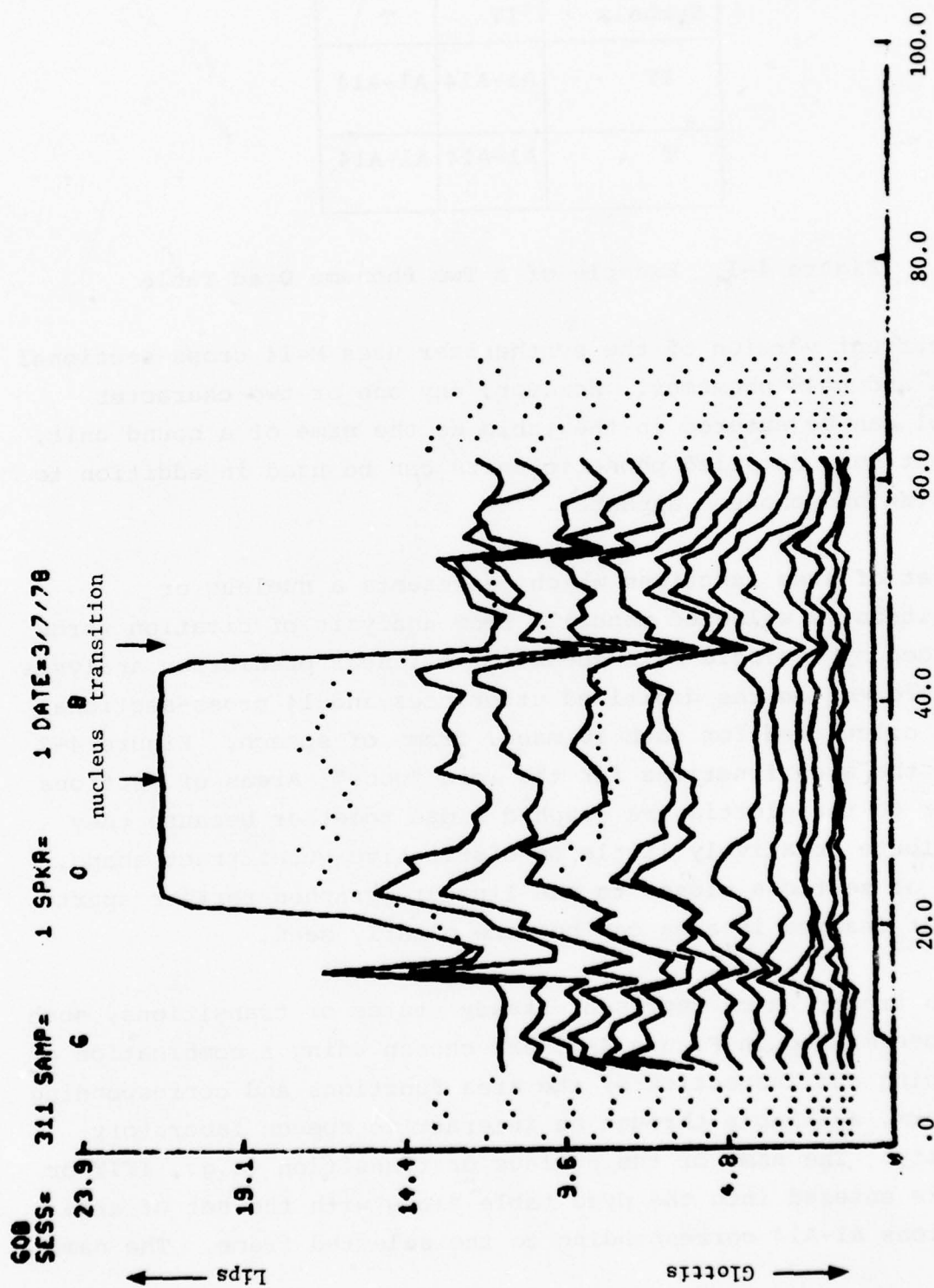
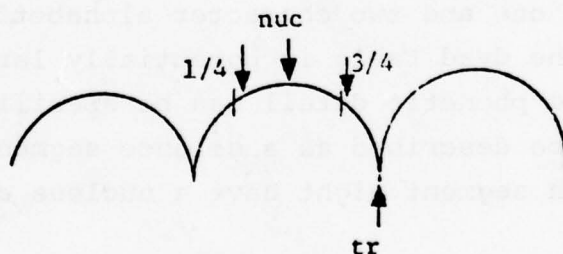


Figure 4-2. Calculated Area Functions from Utterance "gob".

of the utterance from which the speech frame was picked can also be entered for future reference.

In general, steady state vowel nuclei are chosen from contexts which have minimal phonetic influence on the vowel, such as stressed /h_d/ syllables. Single representative steady states of some consonants are almost impossible to characterize, however, as in voiceless aspirated plosives which are made up of both silence and burst portions. To handle such cases, the original dyad table, with a single speech frame representing each nucleus and a single frame representing each transition, was modified to handle more information between frames.

The present version of the dyad table now allows for specification of up to three points (i.e., speech frames) for each transition: the center of the transition, a point midway between the center of the transition and the preceding nucleus, and a point midway between the center of the transition and the following nucleus. Viewed another way, from the point of view of the nucleus, each phoneme in context can be represented as a nucleus, entering 1/4 point, following 3/4 point, and transition:



For synthesis, combinations of these points can be selected in binary according to the following scheme.

<u>1/4 Point</u>	<u>Nucleus</u>	<u>3/4 Point</u>	<u>Transition</u>		
0	0	0	1	=	1
0	0	1	0	=	2
0	0	1	1	=	3
0	1	0	0	=	4
0	1	0	1	=	5
0	1	1	0	=	6
0	1	1	1	=	7
1	0	0	0	=	8
1	0	0	1	=	9
1	0	1	0	=	10
1	0	1	1	=	11
1	1	0	0	=	12
1	1	0	1	=	13
1	1	1	0	=	14
1	1	1	1	=	15

Tests and evaluations presented in this report are based on synthesis version 15, using all points. Informal listening tests of a section of the Rainbow Passage synthesized under all 15 different conditions suggest that version 11 (all transitional information and no nucleus) and version 15 (all transitional information plus nucleus) produce better synthetic speech than the other versions, as might be expected.

It should be noted here that although the terms phoneme, nucleus and transition are being used here to describe the general case, the inventory of one and two character alphabetic symbols that can be used in the dyad table is potentially large enough so that considerable phonetic detail can be specified. For example, a plosive might be described as a silence segment and a burst segment, and each segment might have a nucleus and a 3-point transition.

As part of a phoneme vocoder, the current version of the area function dyad synthesizer receives as input a sequence of triples corresponding to a sequence of phonemes which is the output of the analyzer. The first element of each triple is the phoneme name, one or two alphabetic symbols. The second element of the triple is the duration of the phoneme in 10 msec. segments, and

the third element is a fundamental frequency value in Hertz. An optional fourth element, the LPC gain factor is also available, but is not used in the work reported here. A sample sequence of triples for the word "gob" might be

SI	10	0
G	11	115
AA	27	108
B	17	110
SI	10	0

where SI stands for a silence interval at the beginning and end of the word.

Taking the phonemes pairwise, the synthesizer retrieves appropriate area function parameters from the dyad table, four sets of parameters per dyad. For example, for the dyad GAA, stored parameters for a G nucleus, a GAA transition, and a point on either side of the GAA transition would be retrieved.

A cubic spline interpolation scheme [Akima 1970] is used to generate a full set of area function parameters for the utterance. An appropriate gain contour is generated by rule [Olive 1977]. These parameters can then be used for LPC synthesis.

A graph of the interpolated area functions produced from the transcription given above is shown in Figure 4-3.

4.2 Terminal Analog Synthesis

A digital simulation of terminal analog synthesis [Klatt 1972, 1977] was also tested as a component of the phoneme vocoder. A set of control parameter time functions defined by a sequence of phonemes is used to generate a synthesized speech waveform. Vowels, sonorants and nasals are synthesized by resonators connected in cascade, and other sounds are produced by formant resonators in parallel. The synthesizer includes components

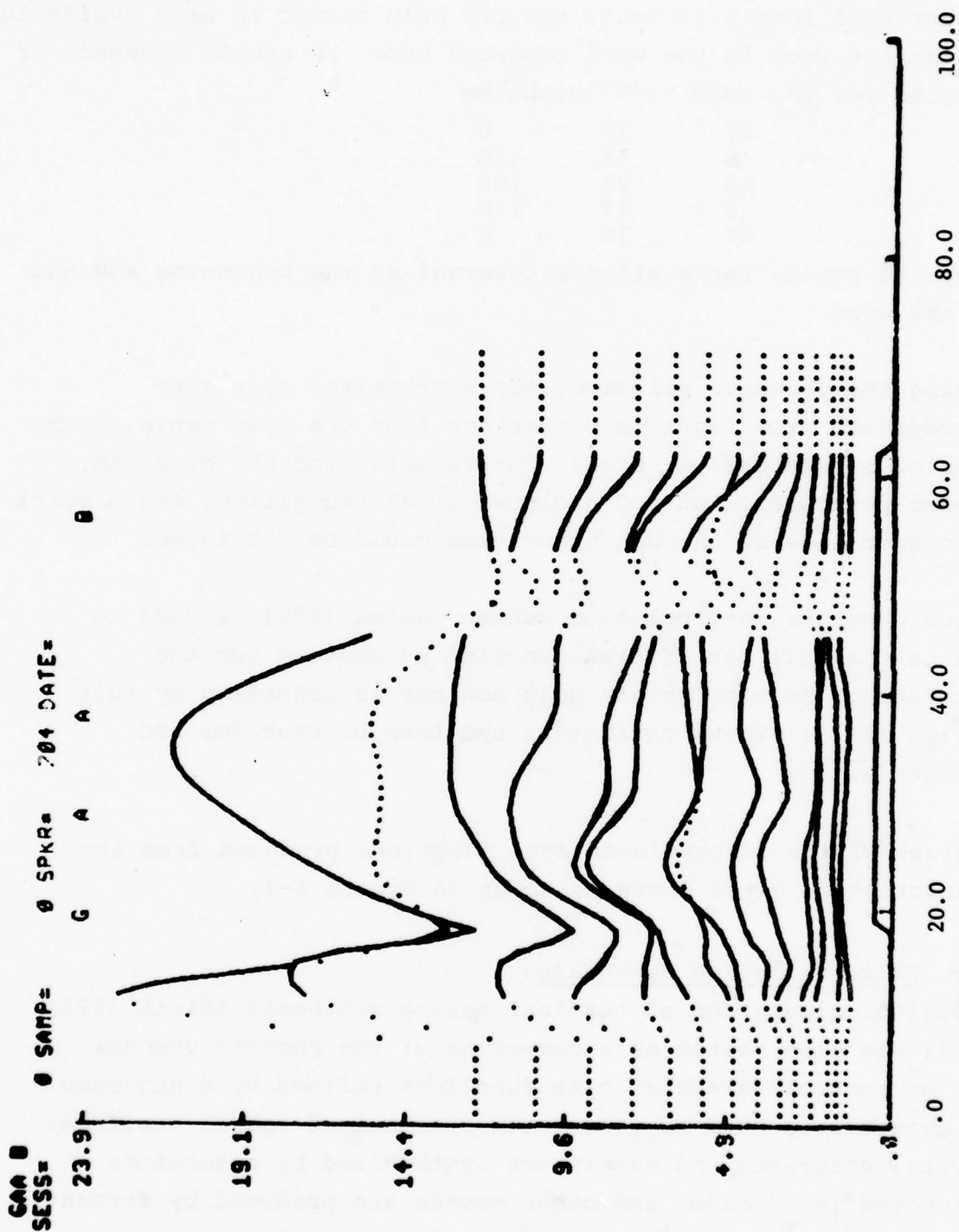


Figure 4-3. Interpolated Area Functions from Transcription of "gob".

which simulate different sound sources, the vocal tract transfer function and sound radiation from the head.

Modifications were made to the existing terminal analog synthesizer to include improvements by Klatt [1977], and to make it compatible with the output of the phoneme vocoder analyzer. Modules generating pitch and duration by rule were removed, as actual pitch and duration is transmitted by the vocoder. Control parameters suggested by Klatt, such as formant bandwidths and nasal zeroes, were added.

Final tuning of the updated synthesis has not been completed, and some work on system interface remains to be done.

5. TESTING

The SDC phoneme vocoder was tested by Dynastat, Inc., using the Diagnostic Rhyme Test (DRT), a two-choice test of consonant discriminability [Voiers, et al. 1973]. The primary purpose of the DRT is to use a panel of listeners to detect specific deficiencies of voice communication systems in transmitting six basic attributes of consonant phonemes: voicing, nasality, sustention, sibilation, graveness, and compactness.

In terms of speech production, the presence or absence of the voicing attribute is determined by whether the vocal folds are vibrating or not, i.e., whether the excitation is periodic or noiselike. The nasality attribute distinguishes sounds made by air passing through the nasal cavity from sounds made by air passing only through the vocal tract. The sustention attribute distinguishes sounds made with sustained passage of air from those made with brief complete closure of the vocal tract. The sibilation attribute distinguishes sounds characterized by strong turbulence and noise-like behavior from those with less turbulence and greater periodicity. The graveness attribute distinguishes sounds articulated at the extreme front of the mouth from those made in a more retracted position. The compactness feature distinguishes sounds articulated toward the back part of the mouth from sounds articulated toward the front. In the taxonomy used in the DRT, combinations of these attributes are sufficient to differentiate among the consonant phonemes of English.

The sets of word pairs which represent the corpus of 96 stimulus words for the DRT test are shown in Table 5-1, grouped according to attribute. With minor exceptions, the initial consonants of the words in each pair differ only in terms of presence vs. absence of the attribute. Twenty experimental words from Dynastat are also included on each test, for a total of 116 items.

Table 5-1. Corpus for DRT Test

Voicing

veal-feel
bean-peen
zoo-sue
dune-tune
gin-chin
dint-tint
vole-foal
goat-coat
zed-said
dense-tense
vault-fault
daunt-taunt
vast-fast
gaff-calf
jock-chock
bond-pond

Nasality

meat-beat
need-deed
moot-boot
news-dues
mitt-bit
rip-dip
moan-bone
note-dote
mend-bend
neck-deck
moss-boss
gnaw-daw
mad-bad
nab-bab
mom-bomb
knock-dock

Sustention

vee-bee
sheet-cheat
foo-pooh
shoes-choose
vill-bill
thick-tick
those-doze
though-dough
then-den
fence-pence
thong-tong
shaw-chaw
than-dan
shad-chad
von-bon
vox-box

Sibilation

zee-thee
cheep-keep
juice-goose
chew-coo
jilt-guilt
sing-thing
joe-go
sole-thole
jest-guest
chair-care
jaws-gauze
saw-thaw
jab-gab
sank-thank
jot-got
chop-cop

Graveness

weed-reed
peak-teak
moon-noon
pool-tool
bid-did
fin-thin
bowl-dole
fore-thor
met-net
pent-tent
fought-thought
bong-dong
bank-dank
fad-thad
wad-rod
pot-tot

Compactness

yield-wield
key-tea
coop-poop
you-rue
hit-fit
gill-dill
ghost-boast
show-so
keg-peg
yen-wren
yawl-wall
caught-taught
gat-bat
shag-sag
hop-fop
got-dot

Dynastat provides analog tapes of male speakers pronouncing single words in the corpus in randomized order. The tapes are processed through the experimental transmission conditions to be tested, and the synthetic speech is submitted to Dynastat's panel of eight listeners. The panel uses score sheets with a

list of word pairs (from Table 5-1) corresponding to the order of the single words on the tape. Each listener hears the synthesized version of each word and decides which item it is in the corresponding word pair.

The SDC phoneme vocoder was tested under four analysis and synthesis conditions.

1. An edited phonemic transcription, produced by manually correcting errors in phoneme classification from the analyzer, was input to the area function dyad synthesizer.
2. The same edited transcription was input to the terminal analog synthesizer.
3. A raw phonemic transcription, produced directly from the analyzer, was input to the area function dyad synthesizer.
4. The same raw transcription was input to the terminal analog synthesizer.

Tapes of two male speakers were used: speaker RH saying List 311B and speaker PK saying List 312A. Test results are discussed in Section 6.

6. EVALUATION

The DRT scores for the four transmission conditions are:

Condition 1 - Edited transcription, area function dyad synthesis	70.6%
Condition 2 - Edited transcription, terminal analog synthesis	83.5%
Condition 3 - Raw transcription, area function dyad synthesis	35.9%
Condition 4 - Raw transcription, terminal analog synthesis	42.8%

The scores given are corrected for effects of chance or guessing by an adjustment of the form

$$P_c = \frac{R-W}{T} (100)$$

where PC is the adjusted percent correct, R is the number of right answers, W the number of wrong answers and T the total number of items involved. A right answer is one which the attribute in the original speech (voicing, nasality, etc.) is perceived to be correctly present or absent in the synthetic speech.

6.1 Condition 1

Figures 6-1 through 6-4 give more detailed analysis of the errors for Conditions 1-4. Figure 6-1a compares the adjusted percent correct scores for the presence and absence of each attribute, and the attribute mean, for Condition 1. Voicing was correctly perceived to be absent in 96.9% of the unvoiced examples, and correctly perceived to be present in only 21.9% of the voiced examples (percentages are given in Figure 6-1c). In other words, under Condition 1, the presence of voicing is not very successfully transmitted. The mean score for voicing is about 60%.

Scores for other attributes for Condition 1 are graphed in Figure 6-1a with actual percentages given in Figure 6-1c. Presence of nasality, presence of

Tracer, Inc.

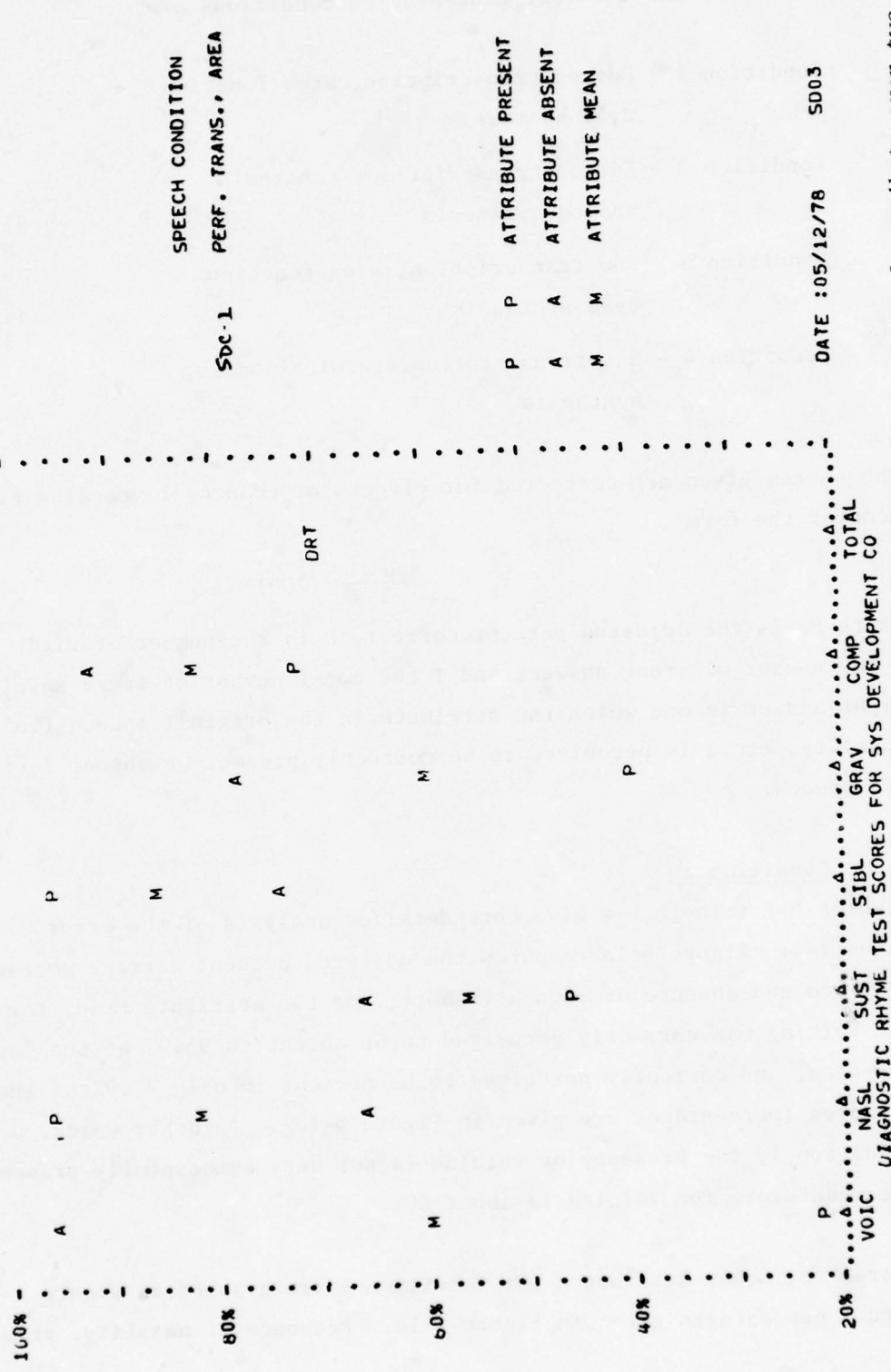


Figure 6-1a. Correct responses (in adjusted percent) for absence and presence of attribute over two speakers, for Condition 1.

Tracor, Inc.

CONTRACTOR : SYS DEVELOPMENT CO										EXPERIMENTAL CONDITION : PERF. TRANS., AREA																			
ERRORS FOR SPEAKERS BY ATTRIBUTES										ERRORS FOR LISTENERS BY ATTRIBUTES																			
ATTRIBUTES PRESENT					ATTRIBUTES ABSENT					ATTRIBUTES TOTAL					ATTRIBUTES TOTAL														
(SPK)	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	(SPK)	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL				
KH	37	1	18	2	16	4	1	78	1	20	12	4	9	5	1	51	KH	36	-19	6	-2	7	-1	0	27				
PK	13	1	16	0	22	14	1	66	1	2	10	13	5	0	0	31	PK	12	-1	6	-13	17	14	1	35				
SUM	50	2	34	2	38	18	2	144	2	22	22	17	14	5	1	82	SUM	48	-20	12	-15	24	13	1	62				
ATTRIBUTES BIAS(SIGN REVERSED)										ATTRIBUTES TOTAL																			
(SPK)	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	(SPK)	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL				
KH	36	-19	6	-2	7	-1	0	27	38	21	30	6	25	9	2	129	KH	36	-19	6	-2	7	-1	0	27				
PK	12	-1	6	-13	17	14	1	35	14	3	26	13	27	14	1	97	PK	12	-1	6	-13	17	14	1	35				
SUM	48	-20	12	-15	24	13	1	62	52	24	56	19	52	23	3	226	SUM	48	-20	12	-15	24	13	1	62				
ERRORS BY ITEMS																													
VC				NS				ST				SB				GV				CM				EX					
ITEMS 1-28				0				4				3				0				7				0					
ITEMS 29-56				2				3				6				0				11				1					
ITEMS 57-84				2				0				4				2				0				0					
ITEMS 85-112				6				0				4				4				0				0					
QUALITY RATINGS										MEAN										S.E.*									
(S) NATURAL										26.2										5.00									
(U) INCONSPICUOUS										71.2										6.34									
(I) INTELLIGIBLE										32.7										4.14									
(T) PLEASANT										39.9										7.14									
(T) ACCEPTABLE										36.6										4.35									
(I) ESTIMATED COMPOSITE										36.4										4.39									
LEVEL : SOFT VS LOUD										49.7										.31									
(S) SPEECH SIGNAL																													
(U) BACKGROUND																													
(I) TOTAL EFFECT																													
STANDARD ERRORS BASED ON LISTENER MEANS																													

* STANDARD ERRORS BASED ON LISTENER MEANS

(S) SPEECH SIGNAL
(U) BACKGROUND
(I) TOTAL EFFECT

Figure 6-1b. Analysis of errors by speaker, attribute and vowel context for Condition 1.

Tracor, Inc.

CONTRACTOR : SYS DEVELOPMENT CO			TEST CONDITION : PERF.		TRANS., AREA		DATE TESTED : 05/12/78	
PRESENT			ABSENT		BIAS		TOTAL	
		S.E.*					S.E.*	S.E.*
VOICING	21.9	4.57	96.9	2.05	-75.0	5.28	59.4	2.36
FRICTIONAL	65.6	9.38	96.9	3.13	-31.3	10.30	81.3	4.72
NONFRICTIONAL	-21.9	7.38	96.9	3.13	-118.7	6.25	37.5	4.72
NASALITY	96.9	2.05	65.6	8.10	31.3	8.52	81.3	4.09
GRAVE	93.8	4.09	34.4	13.31	59.4	14.12	64.1	6.86
ACUTE	100.0	.00	96.9	3.13	3.1	3.13	98.4	1.56
SUSTENTION	46.9	7.38	65.6	6.58	-18.8	11.33	56.3	4.09
VOICED	46.9	11.99	56.3	9.15	-9.4	20.56	51.6	2.83
UNVOICED	46.9	19.15	75.0	8.18	-28.1	22.87	60.9	9.28
SIBILATION	96.9	2.05	73.4	5.51	23.4	6.44	85.2	2.62
VOICED	100.0	.00	87.5	4.72	12.5	4.72	93.8	2.36
UNVOICED	93.8	4.09	59.4	9.38	34.4	10.50	76.6	4.98
GRAVENESS	40.6	4.57	78.1	6.58	-37.5	6.68	59.4	4.57
VOICED	100.0	.00	87.5	6.68	12.5	6.68	93.8	3.34
UNVOICED	-18.8	9.15	68.8	10.30	-87.5	13.36	25.0	7.09
STOPPED	53.1	11.99	71.9	11.02	-18.8	15.49	62.5	8.52
UNSTOPPED	28.1	8.76	84.4	4.57	-56.3	10.30	56.3	4.72
COMPACTNESS	71.9	3.13	92.2	3.29	-20.3	4.05	82.0	2.49
VOICED	96.9	3.13	100.0	.00	-3.1	3.13	98.4	1.56
UNVOICED	46.9	5.66	84.4	6.58	-37.5	9.45	65.6	3.92
SUSTAINED	90.6	4.57	93.8	4.09	-3.1	5.66	92.2	3.29
INTERRUPTED	53.1	5.66	90.6	6.58	-37.5	8.18	71.9	4.57
BK/MJ	53.1	5.66	93.8	4.09	-40.6	6.58	73.4	3.69
BK/FK	90.6	4.57	98.4	5.51	.0	6.68	90.6	3.13
EXPERIMENTAL**	96.9	2.05	98.4	5.51	-1.6	2.83	97.7	1.14

SPEAKER RH PK
LIST # 311B 312A
DRT SCORE 66.4 74.7
S.E.* 1.60 1.74

8 LISTENERS, CREW (01), 192 TOTAL WORDS
2 SPEAKER(S), 96 WORDS PER SPEAKER
STANDARD ERROR FOR SPEAKERS = 4.17
TOTAL VOICED SCORE = 84.4
TOTAL UNVOICED SCORE = 57.0

* STANDARD ERRORS BASED ON LISTENER MEANS.

** EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X TOTAL DRT SCORE = 70.6 X
X STANDARD ERROR = 1.14 X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Figure 6-1c. Correct responses (in adjusted percent) for absence and presence of attribute by sub-type for Condition 1.

Tracor Inc.

CONTRACTOR SYS DEVELOPMENT CO EXPERIMENTAL CONDITION IS PERF. TRANS. AREA
 FOR THE PURPOSES OF FURTHER RESEARCH DESIGNED TO IMPROVE
 YOUR SYSTEM OR DEVICE, YOU WILL FIND IT ADVANTAGEOUS TO GIVE
 SPECIAL ATTENTION TO THE DISTINGUISHABILITY OF THE FOLLOWING WORD PAIRS.*

WORD PAIRS P(C)

45:VOX/BOX **	-37.5
33:FOUGHT/THOUGHT **	-37.5
83:KEY/TEA	-25.0
64:DENSE/TENSE	-25.0
10:SHOES/CHOOSE	-25.0
106:DUNE/TUNE	.0
103:FIN/THIN **	.0
92:GAFF/CALF	.0
15:DINT/TINT	.0
96:PENT/TENT	12.5

** THE CONTRASTS: FAD-THAD, FIN-THIN, FOUGHT-THOUGHT, VON-BON, VOX-BOX, VEE-BEE, VILL-BILL, VAULT-FAULT ARE GENERALLY AMONG THE MOST DIFFICULT TO DISTINGUISH. THEIR PRESENCE ON THE FOREGOING LIST DOES NOT, THEREFORE, REFLECT UNIQUELY UPON THE PERFORMANCE OF YOUR SYSTEM OR DEVICE.

Figure 6-ld. Suggested word pairs for future study for Condition 1.

sibilant, and absence of compactness had relatively high scores. That is, nasals, sibilants and consonants articulated in a relatively front position in the mouth are correctly perceived for Condition 1.

Presence of voicing, as noted before, and presence of sustention and graveness had the lowest scores. That is, voiced consonants are incorrectly perceived as unvoiced, sustained consonants are perceived as not sustained, and labial consonants are perceived as non-labial.

Highest mean scores (over 80%) were obtained for nasality, sibilant and compactness. Lowest mean scores (between 50%-60%) were obtained for voicing, sustention and graveness.

Figure 6-1b breaks down the errors by attribute, speaker and vowel context, and also gives overall quality ratings. In general, more errors were made on transmitted speech of speaker RH (129 total errors) than on speaker PK (97 total errors), although that pattern is not consistent for each attribute. The attribute bias indicates the number and bias (towards presence or absence of the attribute) of the errors for each attribute. For example, for speaker RH, the panel of eight listeners made 37 choices that voicing was absent when it was intended to be present, and one choice that voicing was present when it was intended to be absent. Therefore, the error bias is $37-1=36$ toward perceiving unvoiced sounds when a voiced sound was intended. In the other direction, attribute bias for nasals for speaker RH is $1-20=-19$ toward perceiving a nasal sound when a non-nasal sound is intended.

There appears to be no significant effect of vowel context on perception of presence or absence of the consonant attributes .

Figure 6-1c gives percentage scores for each attribute and also for sub-types for each attribute. For example, the overall score for presence of voicing is 21.9% for Condition 1. However, there is a big difference in the scores for frictional voiced sounds (e.g., /v,z/) vs. non-frictional voiced sounds (e.g., /b,d/). Frictional voiced sounds were perceived as voiced in 65.6% of the cases, and non-frictional voiced sounds were perceived as voiced in

only -21.9% of the cases (negative percentage is possible because of adjusted score; i.e., there were more errors than correct choices).

Figure 6-1d suggests word pairs that represent problem areas such as voicing (dense/tense), sustention (vox/box) and graveness (pent/tent).

6.2 Condition 2

Figure 6-2a gives scores for Condition 2. Mean scores are higher than for Condition 1, and the general pattern is better, although Condition 1 has a higher score for absence of voicing, and a better overall pattern for sibilant and compactness. It is interesting to note that for Condition 2 the correct presence of voicing was perceived more often (95.3%) than correct absence of voicing (67.2), directly opposite to the ranking of voicing scores for Condition 1. This suggests that characteristics from both synthesizers might be used to improve transmission of specific attributes.

Figure 6-2b gives a breakdown of errors for Condition 2. There is no apparent difference in the number of errors for each speaker. The analysis of errors by item is useful because it shows that there were 13 errors for item 13 and 12 errors for item 83, out of a total of 38 errors for the compactness attribute. That is, just two items accounted for two-thirds of the errors for that attribute. However, the same pattern of errors did not occur for Condition 1 so it cannot be attributed to the input transcription.

Vowel context appears to have more influence on consonant attribute scores for Condition 2 than for Condition 1, with differences in vowel corresponding to a range of 66.7% to 93.8% in mean correct scores across attributes.

Figure 6-2c gives the scores by sub-type of attribute. One set of scores which showed considerable variation by sub-type was for absence of voicing, with an overall correct score of 67.2%, but a drop to 34.4% for correct absence of voicing in frictional unvoiced sounds (e.g., /f,s/), and an increase to 100% for correct absence of voicing in non-frictional unvoiced sounds (e.g., /p,t/).

Tracer, Inc.

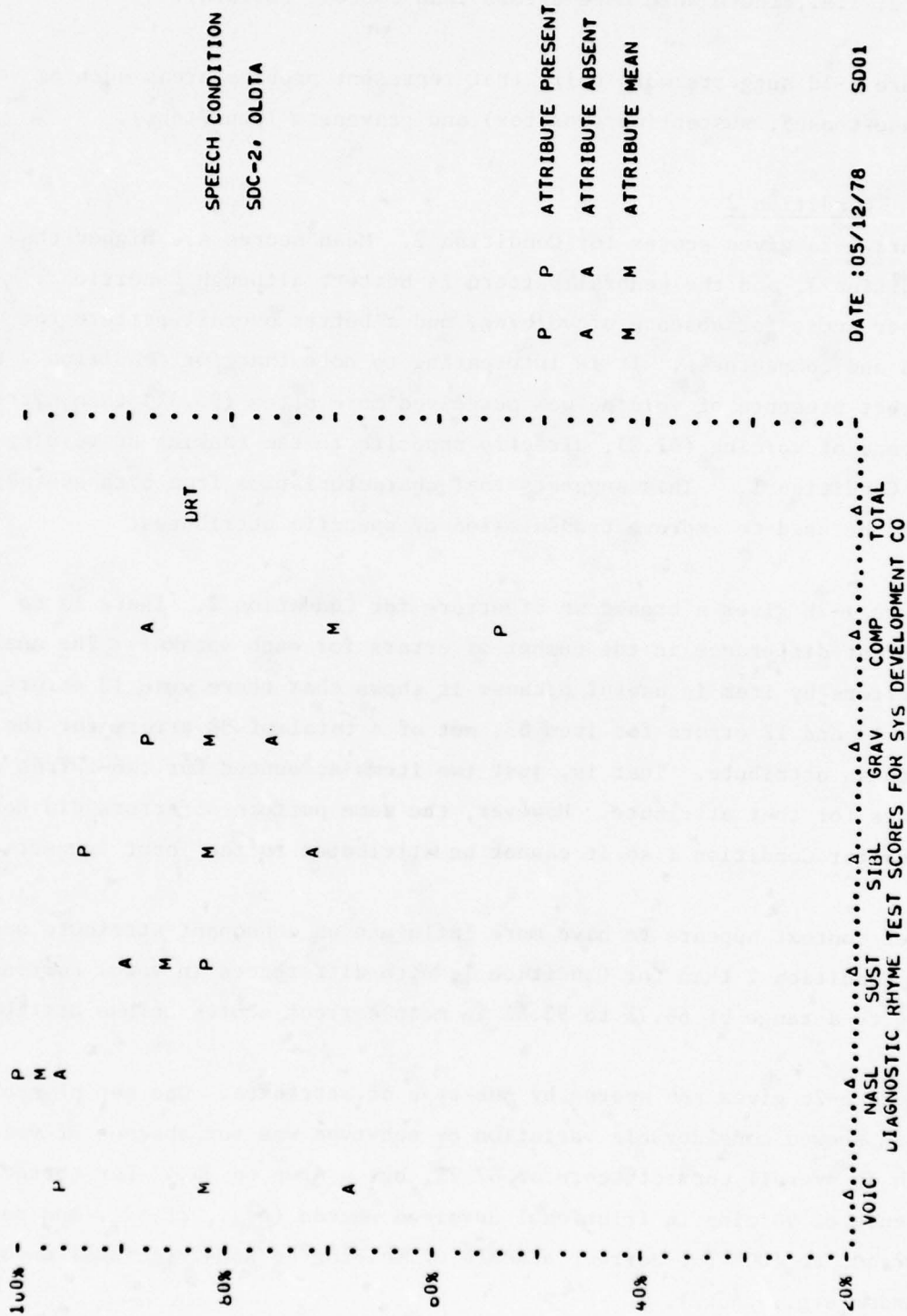


Figure 6-2a. Correct responses (in adjusted percent) for absence and presence of attribute for Condition 2.

Tracor, Inc.

CONTRACTOR : SYS DEVELOPMENT CO										EXPERIMENTAL CONDITION : SUC-2, OLDTA ERRORS FOR SPEAKERS BY ATTRIBUTES													
(SPK)	VOIC	NASL	ATTRIBUTE PRESENT					TOTAL	VOIC	NASL	SUST	ATTRIBUTE ABSENT			EXPL	TOTAL							
			SUST	SIBL	GRAV	COMP	EXPL					SUST	SIBL	GRAV			COMP						
RH	0	0	3	0	8	12	0	23	14	0	1	7	8	0	38								
PK	3	0	8	4	0	18	0	33	7	2	6	11	7	0	33								
SUM	3	0	11	4	8	30	0	56	21	2	7	18	15	8	71								
(SPK)	VOIC	NASL	ATTRIBUTE REVERSED					TOTAL	VOIC	NASL	SUST	ATTRIBUTE TOTAL			EXPL	TOTAL							
			SUST	SIBL	GRAV	COMP	EXPL					SUST	SIBL	GRAV			COMP						
RH	-14	0	2	-7	0	4	0	-15	14	0	4	7	16	20	0	61							
PK	-4	-2	2	-7	-7	18	0	0	10	2	14	15	7	18	0	66							
SUM	-18	-2	4	-14	-7	22	0	-15	24	2	18	22	23	38	0	127							
ERRORS BY ITEMS										VC NS ST SB GV CM EX													
ITEMS 1-28	0	1	0	0	0	1	0	0	3	2	3	13	0	0	0	0	0	0	0				
ITEMS 29-56	0	1	0	2	7	0	0	3	0	1	0	1	6	0	0	0	0	0	0				
ITEMS 57-84	0	0	2	1	0	2	0	0	0	0	0	3	4	0	0	0	0	0	0				
ITEMS 85-112	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2	1				
QUALITY RATINGS										VOWEL CONTEXT										S.E.*			
(S) NATURAL				MEAN			S.E.*													MEAN		S.E.*	
(b) INCONSPICUOUS				35.9			10.15			(EE)										79.2		3.53	
(I) INTELLIGIBLE				76.6			6.07			(IH)										80.2		1.53	
(I) PLEASANT				43.1			6.75			(EH)										93.8		4.10	
(I) ACCEPTABLE				47.0			7.60			(AI)										87.5		2.73	
(I) ESTIMATED COMPOSITE				45.3			6.10			(OO)										92.7		1.04	
				45.1			5.74			(OH)										86.5		2.20	
				50.0			.00			(AW)										81.3		2.62	
				50.0			.00			(AH)										66.7		4.46	
LEVEL : SOFT VS LOUD																							

(S) SPEECH SIGNAL
(B) BACKGROUND
(T) TOTAL EFFECT

* STANDARD ERRORS BASED ON LISTENER MEANS

Figure 6-2b. Analysis of errors by speaker, attribute and vowel context for Condition 2.

CONTRACTOR : SYS DEVELOPMENT CO			TEST CONDITION : SDC-2, OLD TA		DATE TESTED : 05/12/78	
PRESENT	S.E.*	ABSENT	S.E.*	BIAS	TOTAL	S.E.*
VOICING	95.3	67.2	2.29	28.1	81.3	1.67
FRIC TIONAL	90.6	34.4	4.57	56.3	62.5	3.34
NONFRIC TIONAL	100.0	100.0	.00	.0	100.0	.00
NASALITY	100.0	96.9	3.13	3.1	98.4	1.56
GRAVE	100.0	93.8	6.25	6.3	96.9	3.13
ACUTE	100.0	100.0	.00	.0	100.0	.00
SUSTENTATION	82.8	89.1	4.98	-6.3	85.9	2.83
VOICED	71.9	90.6	9.38	-18.8	81.3	5.28
UNVOICED	93.8	87.5	6.68	6.3	90.6	3.13
SIBILATION	93.8	71.9	6.58	21.9	82.8	4.38
VOICED	93.8	68.8	12.27	25.0	81.3	7.83
UNVOICED	93.8	75.0	10.56	18.8	84.4	4.57
GRAVENESS	87.5	76.6	4.98	10.9	82.0	1.84
VOICED	100.0	68.8	7.83	31.3	84.4	3.92
UNVOICED	75.0	84.4	6.58	-9.4	79.7	3.29
STOPPED	100.0	68.8	7.83	31.3	84.4	3.92
UNSTOPPLD	75.0	84.4	6.58	-9.4	79.7	3.29
COMPACTNESS	53.1	87.5	.00	-34.4	70.3	3.69
VOICED	53.1	100.0	.00	-46.9	76.6	4.38
UNVOICED	53.1	75.0	.00	-21.9	64.1	5.51
SUSTAINED	90.6	100.0	.00	-9.4	95.3	2.29
INTERRUPTED	15.6	75.0	.00	-59.4	45.3	6.66
BK/ML	18.8	75.0	.00	-56.3	46.9	6.14
BK/FK	87.5	100.0	.00	-12.5	93.8	3.34
EXPERIMENTAL**	100.0	100.0	6.58	.0	100.0	.00
SPEAKER	RH	PK				
LIST #	311B	312A				
DRT SCORE	84.1	82.8				
S.E.*	1.57	2.42				
8 LISTENERS, CREW (01), 192 TOTAL WORDS						
2 SPEAKER(S), 96 WORDS PER SPEAKER						
STANDARD ERROR FOR SPEAKERS = .65						
TOTAL VOICED SCORE = 80.9						
TOTAL UNVOICED SCORE = 79.7						
* STANDARD ERRORS BASED ON LISTENER MEANS.						
** EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.						

XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X TOTAL DRT SCORE = 83.5 X
X STANDARD ERROR = 1.45 X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Figure 6-2c. Correct responses (in adjusted percent) for absence and presence of attribute by sub-type for Condition 2.

Tracer Inc.

CONTRACTOR SYS DEVELOPMENT CO EXPERIMENTAL CONDITIONS SUC-2, OLDTA

FOR THE PURPOSES OF FURTHER RESEARCH DESIGNED TO IMPROVE
YOUR SYSTEM OR DEVICE, YOU WILL FIND IT ADVANTAGEOUS TO GIVE
SPECIAL ATTENTION TO THE DISTINGUISHABILITY OF THE FOLLOWING WORD PAIRS.*
WORD PAIRS P(C)

71:GIN/CHIN	-62.5
13:GOT/DOT	-62.5
83:KEY/TEA	-50.0
85:JOCK/CHOCK	.0
102:SOLE/THOLE	12.5
45:VOX/BOX **	12.5
33:FOUGT/THOUGHT **	12.5
41:CAUGHT/TAUGHT	25.0
47:8ID/DID	37.5
68:FAD/THAD **	50.0

** THE CONTRASTS: FAD-THAD, FIN-THIN,FOUGHT-THOUGHT,
VON-BON, VOX-BOX, VEE-BEE, VILL-BILL, VAULT-FAULT
ARE GENERALLY AMONG THE MOST DIFFICULT TO DISTINGUISH.
THEIR PRESENCE ON THE FOREGOING LIST DOES NOT, THEREFORE, REFLECT UNIQUELY
UPON, THE PERFORMANCE OF YOUR SYSTEM OR DEVICE.

Figure 6-2d. Suggested word pairs for future study for Condition 2.

Figure 6-2d suggests word pairs which represent problem areas for Condition 2 such as the voicing attribute for friction sounds (gin/chin) and the compactness attribute (key/tea).

6.3 Conditions 3 and 4

The scores for transmission conditions 3 and 4, shown in Figures 6-3 and 6-4, involving raw transcription output from the analyzer which is input to the synthesizer, are much lower than for conditions with edited transcriptions, indicating that errors in analysis are contributing a great deal to the reduced quality of the synthetic speech.

The analysis output is usually over-specific, identifying separate segments which are merged in an edited transcription. For example, two transcriptions of the word "taut" by speaker RH, with phoneme symbol, and durations are:

<u>Raw</u>		<u>Edited</u>	
T	11	T	9
AX	10	AO	21
AA	11	N	5
N	6	T	17
P	10		
S	6		

The analyzer identified two vowels, AX and AA in the same 21 frame (210 msec.) interval where a single vowel AO was labelled in the edited transcription, and identified the aspirated final plosive T as a plosive P followed by considerable friction S.

Further work is needed to smooth the output of the analyzer, and also to include phonetic specificity in the synthesizer which might allow graceful recovery from analysis errors.

Figures 6-3a and 6-4a show the overall pattern of scores for conditions 3 and 4. Scores below 20%, obtained for four attributes for condition 3 and for two attributes for condition 4, do not appear. Correct presence of the

Tracer, Inc.

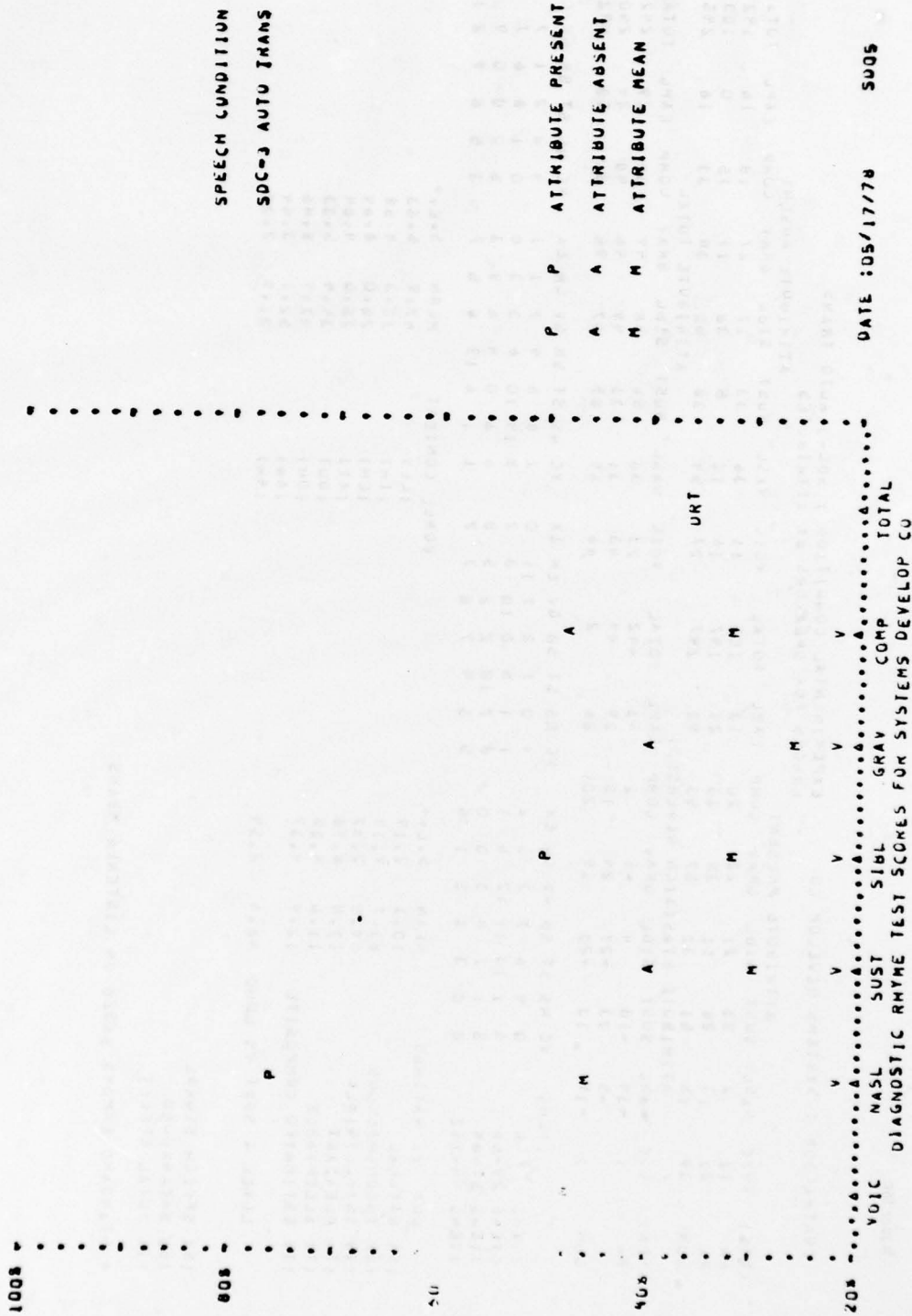


Figure 6-3a. Correct DRT responses (in adjusted percent) for absence and presence of attribute over two speakers for Condition 3.

Tracor, Inc.

CONTRACTOR : SYSTEMS DEVELOP CO EXPERIMENTAL CONDITION : SDC-3 AUTO TRANS
ERRORS FOR SPEAKERS BY ATTRIBUTES

(SPK)	VOIC	NASL	ATTRIBUTE PRESENT			TOTAL	VOIC	NASL	SUST	ATTRIBUTE ABSENT			TOTAL								
			SUST	SIBL	GRAV					COMP	EXPL	SUST		SIBL	GRAV	COMP	EXPL				
RM	12	2	23	21	22	20	13	100	11	36	33	17	27	18	16	142					
PA	27	13	28	11	35	33	29	147	16	18	5	38	11	15	0	103					
SUM	39	15	51	32	57	53	42	247	27	54	38	55	38	33	16	245					
ATTRIBUTE BIAS(SIGN REVERSE)																					
(DFA)	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL					
RM	1	-34	-10	4	-5	2	-3	-42	23	38	56	38	49	38	29	242					
PA	1	-5	23	-27	24	18	29	44	43	31	33	49	46	48	29	250					
SUM	2	-39	13	-23	19	20	26	2	66	69	89	87	95	86	58	492					
LEVEL : SOFT VS LOUD																					
ITEMS																					
VC	NS	ST	SB	GV	CM	EA	VC	NS	ST	SB	GV	CM	EA	VC	NS	ST	SB	GV	CM	EA	
ITEMS 29-56	0	9	6	7	3	3	2	4	0	7	2	7	11	0	2	8	5	4	7	1	1
ITEMS 57-84	8	1	13	11	12	8	1	1	1	5	0	10	8	2	8	14	10	6	3	3	8
ITEMS 85-112	3	1	2	8	3	0	0	8	7	10	2	2	5	8	4	6	0	4	6	9	3
LEVEL : SOFT VS LOUD	8	0	3	8	3	1	5	5	5	4	7	8	7	7	1	1	6	12	6	5	7
QUALITY RATINGS																					
MEAN																					
(S)	NATURAL	INCONSPICUOUS	INTELLIGIBLE	PLEASANT	ACCEPTABLE	ESTIMATED COMPOSITE	MEAN	VOWEL CONTEXT									S.E.*				
VC <td>10.1</td> <td>2.19</td> <td>(EE)</td> <td>47.9</td> <td>5.63</td> <td></td> <td></td> <td colspan="9"></td> <td colspan="3"></td>	10.1	2.19	(EE)	47.9	5.63																
NS	83.7	9.34	(IH)	32.3	4.58																
ST	9.9	2.37	(EM)	25.0	8.64																
SB	17.4	8.78	(AT)	26.0	4.84																
GV	11.6	4.25	(OO)	34.4	5.33																
CM	12.9	4.37	(OM)	47.9	8.46																
EA	46.6	2.54	(AM)	52.1	3.44																
LEVEL : SOFT VS LOUD	46.6	2.54	(AM)	52.1	3.44																

- (S) SPEECH SIGNAL
- (B) BACKGROUND
- (T) TOTAL EFFECT

* STANDARD ERRORS BASED ON LISTENER MEANS

Figure 6-3b. Analysis of errors by speaker, attribute and vowel context for Condition 3.

Tracor Inc.

: SYSTEMS DEVELOP CO			TEST CONDITION : SVC-3 AUTO TRANS		DATE TESTED : 05/17/78	
PRESENT			S.E.	BIAS	S.E.	TOTAL
VOICED	39.1	11.44	57.8	-18.8	14.94	46.4
FRIC TIONAL	75.0	9.45	43.8	31.3	21.52	59.4
ALY- TIONAL	3.1	19.73	71.9	-68.8	23.02	37.5
ASALITY	76.6	7.63	15.6	60.9	10.41	46.1
GRAVE	65.6	10.50	-18.8	84.4	14.89	23.4
ACUTE	67.5	6.68	50.0	37.5	6.68	68.8
SUSTENTION	20.3	5.76	40.6	-20.3	6.50	30.5
VOICED	34.4	12.44	15.6	10.8	16.87	25.0
UNVOICED	6.3	10.30	65.6	-59.4	15.63	38.9
SIBILATION	50.0	8.84	14.1	35.9	11.68	32.0
VOICED	40.6	11.51	15.6	25.0	16.37	28.1
UNVOICED	59.4	8.10	12.5	46.9	9.95	35.9
GRAVENESS	10.9	17.59	40.6	-24.7	19.61	25.8
VOICED	78.1	9.95	-15.6	93.8	19.34	31.3
UNVOICED	-56.3	29.41	96.9	-153.1	27.73	20.3
STOPPED	21.9	13.72	56.3	-34.4	16.32	39.1
UNSTOPPED	.0	22.66	25.0	-25.0	25.44	12.5
COMPATNESS	17.2	6.50	48.4	-31.3	13.98	32.8
VOICED	-3.1	11.99	50.0	-53.1	9.95	23.4
UNVOICED	37.5	8.18	96.9	-9.4	23.59	42.2
SUSTAINED	71.9	5.66	56.3	15.6	11.51	64.1
INTERRUPTED	-37.5	15.67	40.6	-78.1	22.87	1.6
BK/MD	-21.9	12.86	87.5	-109.4	16.32	32.8
BK/FR	56.3	13.15	9.4	46.9	26.07	12.71
EXPERIMENTAL..	34.4	6.44	75.0	-40.6	12.22	54.7
SPEAKER RM PK						
LIST #	3118	312A				
DRT SCORE	37.0	34.9				
S.E.	2.91	4.16				
8 LISTENERS, CKE# (01) 1.192 TOTAL #ORDS						
2 SPEAKER(S), 96 #ORDS PER SPEAKER						
STANDARD ERROR FOR SPEAKERS = 1.04						
TOTAL VOICED SCORE = 27.0						
TOTAL UNVOICED SCORE = 33.6						
* STANDARD ERRORS BASED ON LISTENER MEANS.						
.. EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.						
XX						
X TOTAL DRT SCORE = 36.9 X						
X STANDARD ERROR = 3.02 X						
XX						

Figure 6-3c. Correct responses (in adjusted percent) for absence and presence of attribute by sub-type for Condition 3.

Tracor, Inc.

CONTRACTOR SYSTEMS DEVELOP CO EXPERIMENTAL CONDITIONS SUC-3 AUTO TRANS

FOR THE PURPOSES OF FURTHER RESEARCH DESIGNED TO IMPROVE
YOUR SYSTEM OR DEVICE, YOU WILL FIND IT ADVANTAGEOUS TO GIVE
SPECIAL ATTENTION TO THE DISTINGUISHABILITY OF THE FOLLOWING WORD PAIRS.
WORD PAIRS (PIC)

44:MAD/BAO	-75.0
31:VILL/BILL **	-62.5
111:GILL/DILL	-50.0
102:SOLE/THOLE	-50.0
331FOUGT/THOUGHT **	-50.0
32:JEST/GUEST	-37.5
13:GOT/DOIT	-37.5
66:F00/P00H	-25.0
45:VOX/BOX **	-25.0
40:MET/NET	-25.0

** THE CONTRASTS: FAD-THAD, FIN-TMIN,FOUGHT-THOUGHT,
VON-BON, VOA-BOX, VEE-BEE, VILL-BILL, VAULT-FAULT
ARE GENERALLY AMONG THE MOST DIFFICULT TO DISTINGUISH.
THEIR PRESENCE ON THE FOREGOING LIST DOES NOT, THEREFORE, REFLECT UNIQUELY
UPON THE PERFORMANCE OF YOUR SYSTEM OR DEVICE.

Figure 6-3d. Suggested word pairs for future study for Condition 3.

[illegible]

Figure 6.4a. Correct responses (in adjusted percent) for absence and presence of attribute over two speakers for Condition 4.

CONTRACTOR : SYSTEMS DEVELOP CO										EXPERIMENTAL CONDITION : SOC-4																		
ERRORS FOR SPEAKERS BY ATTRIBUTES										ERRORS FOR SPEAKERS BY ATTRIBUTES																		
ATTRIBUTE PRESENT					ATTRIBUTE ABSENT					ATTRIBUTE PRESENT					ATTRIBUTE ABSENT													
(SPK)	VCIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL				
RH	3	17	15	16	20	9	74	9	9	32	16	27	19	15	6	118												
PK	25	21	7	11	26	49	110	11	16	11	45	34	15	0	137													
SUM	28	38	22	27	46	58	184	25	25	43	32	72	53	30	6	255												
ATTRIBUTE BIAS (SIGN REVERSED)																												
(SPK)	VCIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL <td>VOIC</td> <td>NASL</td> <td>SUST</td> <td>SIBL</td> <td>GRAV</td> <td>COMP</td> <td>EXPL</td> <td>TOTAL <td>VOIC</td> <td>NASL</td> <td>SUST</td> <td>SIBL</td> <td>GRAV</td> <td>COMP</td> <td>EXPL</td> <td>TOTAL</td> </td>	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL <td>VOIC</td> <td>NASL</td> <td>SUST</td> <td>SIBL</td> <td>GRAV</td> <td>COMP</td> <td>EXPL</td> <td>TOTAL</td>	VOIC	NASL	SUST	SIBL	GRAV	COMP	EXPL	TOTAL				
RH	-6	-29	1	-12	-3	5	3	-44	12	35	33	42	35	35	15	192												
PK	9	9	5	-38	-23	11	48	-27	41	31	37	52	45	41	49	247												
SUM	3	-20	6	-50	-26	16	52	-71	53	66	70	94	80	76	64	439												
ERRORS BY ITEMS																												
	VC	NS	ST	SB	GV	CM	EX	VC	NS	ST	SB	GV	CM	EX	VC	NS	ST	SB	GV	CM	EX	VC	NS	ST	SB	GV	CM	EX
ITEMS 1-28	3	1	3	4	0	1	8	0	1	7	7	0	12	0	0	8	4	8	6	0	0	3	8	0	5	2	2	0
ITEMS 29-56	8	7	6	10	12	4	2	4	0	3	0	8	8	1	9	8	12	2	4	5	6	0	0	7	3	0	5	7
ITEMS 57-84	4	0	0	8	4	1	0	1	7	4	4	5	2	14	4	5	1	7	6	5	7	7	5	2	10	7	7	8
ITEMS 85-112	6	0	7	2	14	3	7	2	6	1	11	0	3	0	1	5	7	8	4	7	4	1	5	6	5	8	11	0
QUALITY RATINGS																												
	MEAN	S.E.												MEAN	S.E.													
(S) NATURAL	10.5	2.56												(EE)	47.9	5.85												
(B) INCONSPICUOUS	93.1	2.79												(IM)	32.3	3.32												
(T) INTELLIGIBLE	11.0	2.30												(EM)	50.0	4.73												
(T) PLEASANT	19.9	8.08												(AT)	43.8	6.45												
(T) ACCEPTABLE	13.4	4.03												(OU)	54.2	5.69												
(T) ESTIMATED COMPOSITE	14.8	4.32												(OM)	57.3	4.00												
	43.1	3.56												(AM)	25.0	6.31												
LEVEL : SOFT VS LOUD	43.1	3.56												(AM)	32.3	5.33												

(S) SPEECH SIGNAL
(B) BACKGROUND
(T) TOTAL EFFECT

• STANDARD ERRORS BASED ON LISTENER MEANS

Figure 6-4b. Analysis of errors by speaker, attribute and vowel context for Condition 4.

Tracer, Inc.

CONTRACTOR : SYSTEMS DEVELOP CO			TEST CONDITION : SDC-4		DATE TESTED : 05/19/78	
	PRESENT	S.E..	ABSENT	BIAS	S.E..	TOTAL
VOICING	56.3	5.28	60.9	-4.7	8.82	58.6
FRICTIONAL	62.8	9.45	68.8	-6.3	12.27	65.6
NONFRICTIONAL	50.0	6.68	53.1	-3.1	8.76	51.6
NASALITY	64.1	7.26	32.8	31.3	13.77	48.4
GRAVE	43.8	12.27	18.8	25.0	25.00	31.3
ACUTE	84.4	6.58	46.9	37.5	6.68	65.6
SUSTENTATION	40.6	10.50	50.0	-9.4	14.51	45.3
VOICED	31.3	13.98	50.0	-18.8	19.90	40.6
UNVOICED	50.0	17.03	50.0	.0	23.15	50.0
STABILIZATION	65.6	3.92	-12.5	78.1	10.76	26.6
VOICED	40.6	6.58	-6.3	46.9	15.26	17.2
UNVOICED	90.6	4.57	-18.8	109.4	9.38	35.9
GRAVENESS	57.8	4.69	17.2	40.6	6.58	37.5
VOICED	84.4	4.57	-43.8	128.1	7.38	20.3
UNVOICED	31.3	6.25	78.1	-46.9	9.95	54.7
STOPPED	84.4	4.57	9.4	75.0	9.45	46.9
UNSTOPPED	31.3	9.15	25.0	6.3	15.49	28.1
COMPACTNESS	28.1	5.15	53.1	-25.0	9.74	40.6
VOICED	9.4	12.44	59.4	-50.0	17.68	34.4
UNVOICED	46.9	9.95	46.9	.0	14.94	46.9
SUSTAINED	81.3	7.83	53.1	28.1	17.32	67.2
INTERRUPTED	-25.0	10.56	53.1	-78.1	13.72	14.1
BK/MO	-12.5	11.57	45.6	-78.1	11.99	26.6
BK/FR	68.6	7.83	40.6	28.1	14.51	54.7
EXPERIMENTAL..	9.4	8.10	90.6	-81.3	10.56	50.0
8 LISTENERS, CREW (01), 192 TOTAL WORDS 2 SPEAKER(S), 96 WORDS PER SPEAKER STANDARD ERROR FOR SPEAKERS = 7.16 TOTAL VOICED SCORE = 28.1 TOTAL UNVOICED SCORE = 46.9 * STANDARD ERRORS BASED ON LISTENER MEANS. ** EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.						
XX X TOTAL DRT SCORE = 42.8 X X STANDARD ERROR = 2.45 X XX						

Figure 6-4C. Correct responses (in adjusted percent) for absence and presence of attribute by sub-type for Condition 4.

Tracor, Inc.

CONTRACTOR SYSTEMS DEVELOP CO EXPERIMENTAL CONDITIONS SUC-4

FOR THE PURPOSES OF FURTHER RESEARCH DESIGNED TO IMPROVE
YOUR SYSTEM OR DEVICE, YOU WILL FIND IT ADVANTAGEOUS TO GIVE
SPECIAL ATTENTION TO THE DISTINGUISHABILITY OF THE FOLLOWING WORD PAIRS.
WORD PAIRS PIC)

89:8ONG/DONG	-75.0
45:V0X/BOX **	-50.0
33:FOUGT/THOUGHT **	-50.0
13:GUT/DOOT	-50.0
111:GILL/DILL	-37.5
95:JILT/GILT	-37.5
81:JAWS/GAUZE	-25.0
32:JEST/GUEST	-25.0
43:BEAN/PEEN	-12.5
110:804L/DOOLE	.0

** THE CONTRASTS: FAU-THAD, FIN-THIN, FOUGHT-THOUGHT,
VON-BON, VOR-BOX, VEE-BEE, VILL-BILL, VAULT-FAULT
ARE GENERALLY AMONG THE MOST DIFFICULT TO DISTINGUISH.
THEIR PRESENCE ON THE FOREGOING LIST DOES NOT, THEREFORE, REFLECT UNIQUELY
UPON THE PERFORMANCE OF YOUR SYSTEM OR DEVICE.

Figure 6-4d. Suggested word pairs for future study for Condition 4.

nasality attribute had a relatively high score (76.6%) for condition 3, as did presence of sibilation (65.6%) and nasality (64.1%) for condition 4.

Figures 6-3b and 6-4b give a detailed analysis of the errors for conditions 3 and 4, and Figures 6-3c and 6-4b break down the scores by sub-type of attribute. It is clear that improvements must be made in all areas. Word pairs which represent problem areas are given in Figures 6-3d and 6-4d.

6.4 Diagnostic patterns

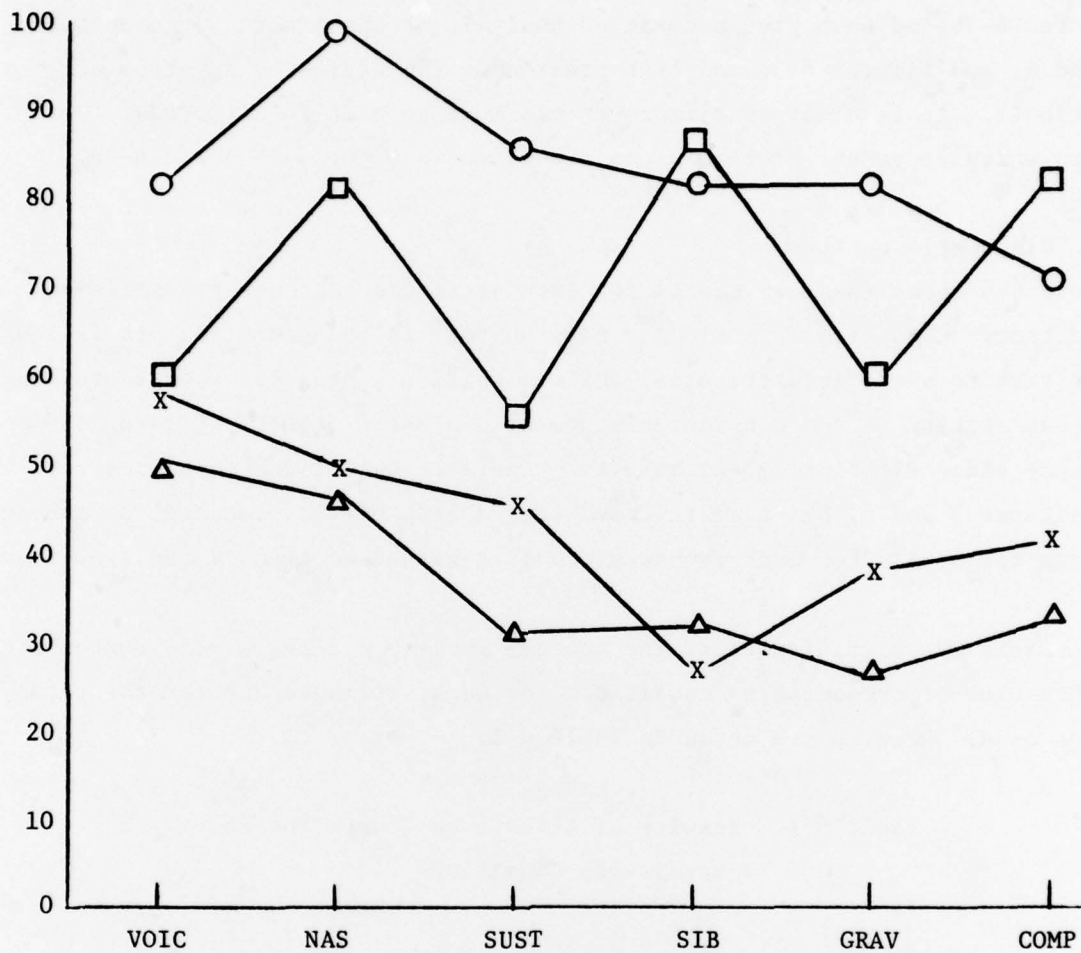
Figure 6-5 shows the mean scores for each attribute for each transmission condition. Condition 2 is clearly more uniform in the sense that it is less sensitive to specific attributes, while condition 1 has very good scores on several attributes and considerably poorer scores on others. Errors in phoneme classification by the analyzer generally depressed the scores for conditions 3 and 4, but also reversed the ranking of the presence vs. absence scores for sibilation and graveness compared with conditions 1 and 2.

To measure the significance of the scores, Student's t-tests were run on each combination of transmission conditions for each attribute and for the total DRT scores. Results are shown in Table 6-1.

Table 6-1. Results of t-tests on Scores for Four Transmission Conditions

<u>Conditions</u>	<u>t for Total Score</u>	
1,2	6.62	P<.001
1,3	10.39	P<.001
1,4	9.63	P<.001
2,3	21.02	P<.001
2,4	13.82	P<.001
3,4	1.96	

That is, as a measure of distinguishing between transmission conditions, the DRT scores are highly significant except for comparison of conditions 3 and 4.



- Condition 1, Edited transcription, area function dyad synthesis
- Condition 2, Edited transcription, terminal analog synthesis
- △ Condition 3, Raw transcription, area function dyad synthesis
- X Condition 4, Raw transcription, terminal analog synthesis

Figure 6-5. DRT Scores for Four Transmission Conditions

7. SUMMARY AND RECOMMENDATIONS

A phoneme vocoder capable of voice transmission at rates lower than 200 bps has been developed which achieved DRT scores of over 80% for two male speakers using terminal analog synthesis, and over 70% using an area function dyad synthesis approach. These scores assumed accurate phoneme transcription as output of the analysis and input to the synthesis. When errorful transcriptions, representing actual analysis output, were used, DRT scores dropped to about 40%.

It is obvious that improvements must be made to the analysis and phoneme classification scheme. One tool that can be used is regression analysis of acoustic parameters to determine their relative contribution to the segmentation and labeling process. A data base exists at SDC of accurately transcribed continuous speech utterances, with each phoneme identified and its time boundaries marked. This corpus could represent ideal values against which to match the regression analysis.

Contextual information should also be incorporated into the final labeling procedure. Currently the only relevant context used by the labeler is the preceding phoneme, but more global context would be very useful. For example, an error in labeling on a DRT item involved the word-final phoneme sequence NPS. By incorporating scores of the phoneme choices (i.e., knowing which choice is most robust), phonotactic constraints and phoneme probabilities, the incorrect sequence could be adjusted to MPS, NTS or NT-aspiration, all of which are possible correct sequences.

Improvements can also be made to the synthesis procedures. The area function dyad synthesizer is designed to be flexible enough to handle phonetic specificity, so that sound units other than phoneme level units, such as the burst components of stops or the off-glides of diphthongs, can be entered into the dyad table. Interpolation over such points should more closely approximate actual speech than current interpolation over phoneme points only.

Modifications to the terminal analog synthesizer are also available and allow more variable acoustic parameters, such as nasal zeroes and formant bandwidths,

than the version used in the current study. The resulting synthetic speech waveforms should more closely represent actual speech waveforms because the production mechanism is more accurately modeled.

The development of the SDC phoneme vocoder has demonstrated that a low data rate voice transmission system is available that yields reasonably high scores on standardized tests such as the DRT. However, substantial improvements remain to be made to both analysis and synthesis.

8. REFERENCES

- Akima, H., A new method of interpolating and smooth curve fitting based on local procedures, J. Assoc. Comp. Mach. 589-602 (1970).
- Atal, B. S., Determination of the vocal tract shape directly from the speech wave, J. Acoust. Soc. Am. 47, 65(A) (1970).
- Atal, B.S. and S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am. 50, 637-655 (1971).
- Atal B. S. and M. R. Schroeder, Predictive coding of speech signals, Proc. 1967 Conf. Speech Commun. and Process, 360-361 (1967).
- Bernstein, M. I., Interactive systems research: final report to the director, Advanced Research Projects Agency, SDC Tech. Memo-5243/004/00, November 1975.
- Dudley, H., The vocoder, Bell Lab. Record 17, 122-126 (1939).
- Flanagan, J. L., Speech Analysis Synthesis and Perception, Springer-Verlag (1965).
- Flanagan, J. L., C. H. Coker and C. M. Bird, Computer simulation of a formant-vocoder synthesizer, J. Acoust. Soc. Am. 34, 2003 (A) (1962).
- Gillmann, R. A., A fast frequency domain pitch algorithm, J. Acoust. Soc. Am. 58, Supplement No. 1, S62 (1975).
- Gold, B., Digital speech networks, Proc. IEEE 65, 1636-1658 (1977).
- Gold, B. and C. M. Rader, The channel vocoder, IEEE Trans. Audio Electroacoustics AU-15, 148-160 (1967).
- Itakura, F. and S. Saito, Analysis synthesis telephony based upon the maximum likelihood method, Reports of the 6th Int. Cong. Acoustics, C-5-5, Tokyo (1968).
- Kameny, I., Automatic acoustic-phonetic analysis of vowels and sonorants, Conf. Record IEEE Int. Conf. Acoustics, Speech, Signal Proc., 166-169 (1976).
- Kameny, I., W. A. Brackenridge and R. Gillmann, Automatic formant tracking, J. Acoust. Soc. Am. 56, Supplement No. 1, S28 (1974).
- Klatt, D. H., Acoustic Theory of terminal analog speech synthesis, Proc. 1972 Int. Conf. Speech Comm. and Proc., IEEE Catalog No. 72, CH 0567-7 AE, 131-135 (1972).
- Klatt, D. H., Acascade/parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis, J. Acoust. Soc. Am. 61, Supplement No. 1, S68 (1977).
- Makhoul, J., Linear prediction: a tutorial review, Proc. IEEE 63, 561-580 (1975).

Markel, J. D., Formant trajectory estimation from a linear least-squares inverse filter formulation, SCRL Monograph No. 7, Speech Communications Research Laboratory, Santa Barbara, California (1971).

Markel, J. D. and A. H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag (1976).

Markel, J. D., A. H. Gray and H. Wakita, Linear prediction of speech-theory and practice, SCRL Monograph No. 10, Speech Communications Research Laboratory, Santa Barbara, California (1973).

Mermelstein, P., Automatic segmentation of speech into syllabic units, J. Acoust. Soc. Am. 58, 880-883 (1975).

Molho, L., Automatic acoustic phonetic analysis of fricatives and plosives, Conf. Record IEEE Int. Conf. Acoustics, Speech Signal Proc., 182-185 (1976).

Olive, J. P., Rule synthesis of speech from dyadic units, Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc., 568-570 (1977).

Oppenheim, A. V., Speech analysis-synthesis system based on homomorphic filtering, J. Acoust. Soc. Am. 45, 459-462 (1969).

Ritea, H. B., Automatic speech understanding systems, Proc. 11th IEEE Computer Soc. Conf., Washington, D.C. (1975).

Saito, S. and F. Itakura, The theoretical consideration of statistically optimum methods for speech spectral density, Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo (1966) (in Japanese).

Schroeder, M. R., Vocoders: analysis and synthesis of speech, Proc. IEEE 54, 720-734 (1966).

Skinner, T. E., Autocorrelation method of determining fundamental frequency, Univac Intercommunication, April 1973.

Voiers, W. D., A. D. Sharpley and C. H. Hehmsoth, Research on diagnostic evaluation of speech intelligibility, Final Report, AFSC Contract No. F19628-70-C-0182 (1973).

Wakita, H., Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods, SCRL Monograph No. 9, Speech Communications Research Laboratory, Santa Barbara, California (1972).

Weinstein, C. J., S. S. McCandless, L. F. Mondschein, and V.W. Zue, A system for acoustic-phonetic analysis of continuous speech, IEEE Trans. Acoustic Speech, Signal Proc. ASSP-23, 54-67 (1975).